P-Values for Classification – Computational Aspects and Asymptotics

Inauguraldissertation
der Philosophisch-naturwissenschaftlichen Fakultät
der Universität Bern
und
der Fakultät für Mathematik und Informatik
der Georg-August-Universität Göttingen

vorgelegt von

Niki Roger Zumbrunnen
von Aeschi bei Spiez

Leiter der Arbeit:
Prof. Dr. L. Dümbgen
Institut für mathematische Statistik und Versicherungslehre
der Universität Bern
und
Prof. Dr. A. Munk
Institut für mathematische Stochastik
der Georg-August-Universität Göttingen

P-Values for Classification – Computational Aspects and Asymptotics

Inauguraldissertation

der Philosophisch-naturwissenschaftlichen Fakultät

der Universität Bern

und

der Fakultät für Mathematik und Informatik

der Georg-August-Universität Göttingen

vorgelegt von

Niki Roger Zumbrunnen

von Aeschi bei Spiez

Leiter der Arbeit:

Prof. Dr. L. Dümbgen

Institut für mathematische Statistik und Versicherungslehre

der Universität Bern

und

Prof. Dr. A. Munk

Institut für mathematische Stochastik

der Georg-August-Universität Göttingen

Von der Philosophisch-naturwissenschaftlichen Fakultät angenommen.

Bern, 05. 03. 2014

Der Dekan:

Prof. Dr. S. Decurtins

Abstract

P-Values for Classification. Let $(\boldsymbol{X}, Y)$ be a random variable consisting of an observed feature vector $\boldsymbol{X}$ and an unobserved class label $Y = 1, 2, \ldots, L$ with unknown joint distribution. In addition, let $\mathcal{D}$ be a training data set consisting of n completely observed independent copies of $(\boldsymbol{X}, Y)$.

First, we consider a point predictor for Y, namely the standard linear classifier for two classes. But we do not assume Gaussian distributions. In this setting we provide a central limit theorem for missclassification rates and cross-validated estimators thereof.

Point predictors do not provide information about confidence. To get such information, we construct for each $b = 1, 2, \ldots, L$ a p-value $\pi_b(\boldsymbol{X}, \mathcal{D})$ for the null hypothesis that $Y = b$, treating Y temporarily as a fixed parameter, i.e. we construct a prediction region for Y with a certain confidence. In particular, we consider p-values based on the plug-in statistic for the standard model with two classes and prove a central limit theorem for inclusion probabilities and cross-validated estimators thereof.

In addition, we discuss data-driven choices of tuning parameters for p-values based on multicategory logistic regression, where we use regularization terms to deal with high-dimensional feature vectors $\boldsymbol{X}$.

Randomizded P-Values. Randomized tests are a familiar concept from mathematical statistics. The goal is to obtain tests with exact prescribed significance level even in settings with test statistics having discrete distributions. We discuss the related concept of randomized p-values. One benefit is that p-values obtained from different independent test statistics can be combined more easily. Since in applications non-randomized tests and p-values are needed, we review and modify a method of Meinshausen et al. (2009).

A major example is the analysis of several independent contingency tables with small cell counts. We propose various ways of combining corresponding randomized p-values. We also illustrate the benefits of the final de-randomized test.

Acknowledgements

First and foremost, I would like to thank the main supervisor, Lutz Dümbgen, for his excellent support throughout my studies. His always-open office door, his positive attitude and his patience made it a great pleasure to work with him.

A special thanks goes to the co-supervisor, Axel Munk, for his valuable inputs and for giving me the opportunity of spending a month at the University of Göttingen. I am also grateful to Prof. Dr. Regina Liu for reviewing my thesis.

Many thanks also to Jasmin Wandel, Dominic and Heike Schuhmacher, Benjamin Baumgartner, David Ginsbourger, Chris Kopp, Bernhard Freiermuth and all other members of the IMSV for helpful and interesting discussions, table football and other enjoyable breaks.

Finally, I would like to thank my family and Olivia Jutzi for their support and encouragement.

This work was supported by the research group FOR916 of the Swiss National Science Foundation (SNF) and the Deutsche Forschungsgemeinschaft (DFG).

Contents

Overview

We start in Section 1.1 with a short introduction to classification. In the remaining part of Chapter 1 we present the p-values for classification introduced by Dümbgen et al. (2008). First we assume that the joint distribution of $(\boldsymbol{X}, Y)$ is known. In this setting optimal p-values are available. If the joint distribution is unknown, we use training data to compute nonparametric p-values based on permutation tests. We review asymptotic results of Dümbgen et al. (2008) and Zumbrunnen (2009). Finally, we comment on technical details and the implementation of the p-values in the R-package `pvclass`.

Some of the test statistics we use depend on a tuning parameter such as the k in the nearest neighbor method or the penalty parameter τ in the logistic regression. In Chapter 2 we propose a data-driven way to choose such parameters. In addition, we comment on computational issues.

The theoretic main result is given in Chapter 3. First we consider linear discriminant analysis with two classes. But we do not assume Gaussian distributions. To estimate the covariance matrix we use either the standard estimator or more robust M-estimators. In these two settings we present central limit theorems for missclassification rates and cross-validated estimators thereof. This result implies in particular that the estimators are root-n-consistent.

Next we consider p-values based on the plug-in statistic for the standard model with two classes. But we relax the assumption of Gaussian distributions to elliptical symmetry. The corresponding conditional inclusion probabilities are of interest to judge the separability of the two classes. However, these theoretic quantities are typically unknown. Therefore we use cross-validation to estimate them. Dümbgen et al. (2008) proved that these estimators are consistent. We take a closer look at the inclusion probabilities and the corresponding estimators and describe their asymptotic distribution. In particular, we derive a central limit theorem, which implies that the estimators are root-n-consistent. Moreover, it enables us to construct confidence intervals for the inclusion probabilities.

For the computation of the p-values, we add the new observation temporarily to a certain class. But it may be an outlier with respect to the distribution of this class. Therefore it is reasonable to use a robust M-estimator for the covariance matrix. Our asymptotic results are valid for the standard estimator as well as for the M-estimators.

Chapter 4 is not directly related to the main part of this thesis. In this chapter we discuss the concept of randomized p-values. Since in applications non-randomized tests and p-values are needed, we review and modify a method of Meinshausen et al. (2009). We propose various ways of combining corresponding randomized p-values. We also illustrate the benefits of the final de-randomized test for several independent contingency tables with small cell counts.

1. Classifiers and P-Values

In this chapter we first give a short introduction to classification. For a more detailed introduction and further references we refer to McLachlan (1992). Then we introduce nonparametric p-values for classification as they are given in Dümbgen et al. (2008) and comment on the implementation in the `R`-package `pvclass`.

This chapter is mainly based on Zumbrunnen (2009) and Dümbgen (2011).

1.1. Classification

Let $(\boldsymbol{X}, Y)$ be a pair of random variables, where

$$Y \in \mathcal{Y} := \{1, \ldots, L\}, \quad L \geq 2$$

is the *class label* of an observation, which is described by the *feature vector* $\boldsymbol{X}$ with values in the *feature space* $\mathcal{X}$.

Classifying $(\boldsymbol{X}, Y)$ means that only $\boldsymbol{X}$ is observed and Y has to be predicted via $\boldsymbol{X}$.

1.1.1. Optimal Classifiers in the Ideal Case

Suppose that the joint distribution of $(\boldsymbol{X}, Y)$ is known, i.e. we know the *prior probabilities*

$$w_\theta := \mathbb{P}(Y = \theta)$$

and the conditional distributions

$$P_\theta := \mathcal{L}(\boldsymbol{X} \mid Y = \theta)$$

for all $\theta \in \mathcal{Y}$. Further let M be a measure on $\mathcal{X}$ and let each conditional distribution P_θ be described by a density function f_θ with respect to M, i.e.

$$P_\theta(B) = \int_B f_\theta(\boldsymbol{x})\, M(\mathrm{d}\boldsymbol{x}),$$

for measurable sets $B \subset \mathcal{X}$.

In the simplest case a *classifier* is a point predictor $\widehat{Y}(\boldsymbol{X})\colon \mathcal{X} \to \mathcal{Y}$ for Y. To find an optimal classifier we need a quality criterion, for example the risk of missclassification

$$R(\widehat{Y}) := \mathbb{P}(\widehat{Y}(\boldsymbol{X}) \neq Y).$$

The following proposition characterizes *optimal classifiers* $\widehat{Y}^*(\boldsymbol{X})$, in the sense that they minimize the risk R:

Lemma 1.1. *Let $\widehat{Y}(\boldsymbol{X})\colon \mathcal{X} \to \mathcal{Y}$ be a classifier. Then*

$$R(\widehat{Y}) \geq 1 - \int \max_{\theta \in \mathcal{Y}} w_\theta f_\theta(\boldsymbol{x})\, M(\mathrm{d}\boldsymbol{x}),$$

with equality if and only if

$$\widehat{Y}(\boldsymbol{x}) \in \arg\max_{\theta \in \mathcal{Y}} w_\theta f_\theta(\boldsymbol{x}) \quad \textit{for } M\textit{-almost all } \boldsymbol{x} \in \mathcal{X}. \tag{1.1}$$

Proof. With $\mathbb{1}(A)$ denoting the indicator function for the set A,

$$\begin{aligned} R(\widehat{Y}) &:= \mathbb{P}(\widehat{Y}(\boldsymbol{X}) \neq Y) \\ &= 1 - \sum_{\theta \in \mathcal{Y}} w_\theta \mathbb{P}(\widehat{Y}(\boldsymbol{X}) = \theta \mid Y = \theta) \\ &= 1 - \sum_{\theta \in \mathcal{Y}} w_\theta \int f_\theta(\boldsymbol{x}) \mathbb{1}\{\widehat{Y}(\boldsymbol{x}) = \theta\}\, M(\mathrm{d}\boldsymbol{x}) \\ &= 1 - \int w_{\widehat{Y}(\boldsymbol{x})} f_{\widehat{Y}(\boldsymbol{x})}(\boldsymbol{x})\, M(\mathrm{d}\boldsymbol{x}) \\ &\geq 1 - \int \max_{\theta \in \mathcal{Y}} w_\theta f_\theta(\boldsymbol{x})\, M(\mathrm{d}\boldsymbol{x}). \end{aligned}$$

The inequality is obviously an equality if and only if (1.1) is satisfied. □

Standard Gaussian Model

Let $P_\theta = \mathcal{N}_d(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma})$ with mean vectors $\boldsymbol{\mu}_\theta \in \mathbb{R}^d$ and a common symmetric, positive definite covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$.

We may write the Gaussian densities as

$$f_\theta(\boldsymbol{x}) = c \exp\big(-D^2_{\boldsymbol{\Sigma}}(\boldsymbol{x}, \boldsymbol{\mu}_\theta)/2\big)$$

with $c := (2\pi)^{-d/2} \det(\boldsymbol{\Sigma})^{-1/2}$ and the *Mahalanobis distance*

$$D_{\boldsymbol{\Sigma}}(\boldsymbol{x}, \boldsymbol{y}) := \sqrt{(\boldsymbol{x} - \boldsymbol{y})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{y})} = \|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{y})\|$$

between $\boldsymbol{x} \in \mathbb{R}^d$ and $\boldsymbol{y} \in \mathbb{R}^d$ (with respect to $\boldsymbol{\Sigma}$). Here and throughout $\|\cdot\|$ denotes the Euclidean norm for vectors or the Frobenius norm for matrices, respectively.

Therefore the optimal classifier can be characterized by

$$\widehat{Y}^*(\boldsymbol{x}) \in \operatorname*{arg\,min}_{\theta \in \mathcal{Y}} \big(D^2_{\boldsymbol{\Sigma}}(\boldsymbol{x}, \boldsymbol{\mu}_\theta) - 2 \log w_\theta \big). \tag{1.2}$$

Suppose that $L = 2$ and $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$. Then

$$D^2_{\boldsymbol{\Sigma}}(\boldsymbol{x}, \boldsymbol{\mu}_1) - 2 \log w_1 \left\{ \begin{matrix} > \\ = \\ < \end{matrix} \right\} D^2_{\boldsymbol{\Sigma}}(\boldsymbol{x}, \boldsymbol{\mu}_2) - 2 \log w_2$$

if and only if

$$(\boldsymbol{x} - \boldsymbol{\mu}_{1,2})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \left\{ \begin{matrix} > \\ = \\ < \end{matrix} \right\} \log(w_1 / w_2),$$

with $\boldsymbol{\mu}_{\theta,b} := (\boldsymbol{\mu}_\theta + \boldsymbol{\mu}_b)/2$ for $\theta, b \in \mathcal{Y}$. The sets of all feature vectors assigned uniquely to class 1 or 2, respectively, are separated by a hyperplane orthogonal to $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$.

1.1.2. Classification Using Training Data

The joint distribution of $(\boldsymbol{X}, Y)$ is typically unknown. In this case we estimate the prior probabilities w_θ and the densities f_θ by adequate estimators $\widehat{w}_\theta(\mathcal{D})$ and $\widehat{f}_\theta(\cdot, \mathcal{D})$, respectively. Then we choose a classifier

$$\widehat{Y}(\boldsymbol{x}, \mathcal{D}) \in \operatorname*{arg\,max}_{\theta \in \mathcal{Y}} \widehat{w}_\theta \widehat{f}_\theta(\boldsymbol{x}).$$

To estimate w_θ and f_θ we use *training data* $\mathcal{D}$, consisting of pairs $(\boldsymbol{X}_i, Y_i)$, for $i = 1, \ldots, n$, whereas the $\boldsymbol{X}_i$ as well as the Y_i are known. We consider the $(\boldsymbol{X}_i, Y_i)$ as random variables with the same distribution as $(\boldsymbol{X}, Y)$, and assume that the $n+1$ data pairs $(\boldsymbol{X}, Y), (\boldsymbol{X}_1, Y_1), (\boldsymbol{X}_2, Y_2), \ldots, (\boldsymbol{X}_n, Y_n)$ are stochastically independent. Let

$$\mathcal{G}_\theta := \{ i \leq n \colon Y_i = \theta \} \quad \text{and} \quad N_\theta := \#\mathcal{G}_\theta.$$

Then an estimator for w_θ is given by

$$\widehat{w}_\theta := \frac{N_\theta}{n}.$$

Linear Discriminant Analysis

In the standard model with $P_\theta = \mathcal{N}_d(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma})$ we replace the unknown mean vectors $\boldsymbol{\mu}_\theta$ and covariance matrix $\boldsymbol{\Sigma}$ in (1.2) with corresponding estimators and get the *standard linear classifier*

$$\widehat{Y}^*(\boldsymbol{x}) \in \operatorname*{arg\,min}_{\theta \in \mathcal{Y}} \big(D^2_{\widehat{\boldsymbol{\Sigma}}}(\boldsymbol{x}, \widehat{\boldsymbol{\mu}}_\theta) - 2 \log \widehat{w}_\theta \big). \tag{1.3}$$

The standard estimators in this model are given by

$$\widehat{\boldsymbol{\mu}}_\theta := \frac{1}{N_\theta} \sum_{i \in \mathcal{G}_\theta} \boldsymbol{X}_i,$$
$$\widehat{\boldsymbol{\Sigma}} := \frac{1}{n - L} \sum_{\theta \in \mathcal{Y}} \sum_{i \in \mathcal{G}_\theta} (\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_\theta)(\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_\theta)^\top.$$

The assumption of Gaussian distributions could be relaxed, e.g. to elliptically symmetric distributions.

Elliptically Symmetric Distributions. The random vector $\boldsymbol{Z} \in \mathbb{R}^d$ has a *spherically symmetric* distribution (with respect to $\mathbf{0}$) if for any orthonormal matrix $\boldsymbol{B} \in \mathbb{R}^{d \times d}$, $\boldsymbol{BZ}$ has the same distribution as $\boldsymbol{Z}$. The distribution of a random vector $\boldsymbol{X} \in \mathbb{R}^d$ is *elliptically symmetric* with center $\boldsymbol{\mu} \in \mathbb{R}^d$ and scatter matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$, if $\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu})$ has a spherically symmetric distribution.

For a spherically symmetric random vector $\boldsymbol{Z} \in \mathbb{R}^d$ with $\mathbb{P}(\boldsymbol{Z} = \mathbf{0}) = 0$ and any unit vector $\boldsymbol{v} \in \mathbb{R}^d$

$$\mathcal{L}(\boldsymbol{v}^\top \boldsymbol{Z}) = \mathcal{L}(Z_1), \tag{1.4}$$

where Z_1 is the first component of $\boldsymbol{Z}$. For the proof of this claim and more details on elliptically symmetric distributions we refer to Muirhead (1982).

Robust M-Estimators. The standard estimator for $\boldsymbol{\Sigma}$ is sensitive to outliers. As an alternative, we consider the more robust M-estimators $\widehat{\boldsymbol{\Sigma}}_M$ and $\widehat{\boldsymbol{\Sigma}}_{sym}$. The first M-estimator, $\widehat{\boldsymbol{\Sigma}}_M$, is the maximum likelihood estimator in the model where $P_\theta = \mathcal{N}_d(\boldsymbol{\mu}_\theta, c_\theta \boldsymbol{\Sigma})$ with $c_\theta > 0$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ symmetric and positive definite with $\det(\boldsymbol{\Sigma}) = 1$. For the calculations we use that $\widehat{\boldsymbol{\Sigma}}_M$ is the solution of the fixed point equation

$$\boldsymbol{\Sigma} = d \sum_{\theta=1}^{L} \frac{N_\theta}{n} \frac{\boldsymbol{M}_\theta}{\operatorname{trace}(\boldsymbol{\Sigma}^{-1} \boldsymbol{M}_\theta)}$$

with $\boldsymbol{M}_\theta := \sum_{i \in \mathcal{G}_\theta} (\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_\theta)(\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_\theta)^\top$.

The second M-estimator, $\widehat{\boldsymbol{\Sigma}}_{sym}$, is a generalization for more than one group of the symmetrized version of Tyler's M-estimator, as it is defined in Dümbgen (1998). We assume that the observations $\boldsymbol{X}_i$ are pairwise different within groups. Then $\widehat{\boldsymbol{\Sigma}}_{sym}$ is the solution of the fixed point equation

$$\boldsymbol{\Sigma} = \frac{d}{c} \sum_{\theta=1}^{L} \frac{1}{N_\theta} \sum_{\substack{i,j \in \mathcal{G}_\theta \\ i<j}} \frac{(\boldsymbol{X}_i - \boldsymbol{X}_j)(\boldsymbol{X}_i - \boldsymbol{X}_j)^\top}{(\boldsymbol{X}_i - \boldsymbol{X}_j)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{X}_i - \boldsymbol{X}_j)}$$

with $c := \sum_{\theta=1}^{L} (N_\theta - 1)/2 = (n - L)/2$.

k Nearest Neighbors

The nearest-neighbor method is a nonparametric alternative to estimate f_θ. It has the advantage, that only few assumptions about the distributions P_θ are required. Suppose that $(\mathcal{X}, d)$ is a separable metric space and consider the closed balls

$$B(\boldsymbol{x}, r) := \{\boldsymbol{y} \in \mathcal{X} \colon d(\boldsymbol{x}, \boldsymbol{y}) \le r\}$$

and the open balls

$$U(\boldsymbol{x}, r) := \{\boldsymbol{y} \in \mathcal{X} \colon d(\boldsymbol{x}, \boldsymbol{y}) < r\}$$

for all $\boldsymbol{x} \in \mathcal{X}$ and $r \ge 0$. Assume that

$$M\big(B(\boldsymbol{x}, r)\big) < \infty \quad \text{for all } \boldsymbol{x} \in \mathcal{X} \text{ and } r \ge 0$$

and that f_θ is continuous for all $\theta \in \mathcal{Y}$.

Lemma 1.2. *Let $\mathcal{X}_0$ be the support of $\mathcal{L}(\boldsymbol{X})$, i.e.*

$$\mathcal{X}_0 := \{\boldsymbol{x} \in \mathcal{X} \colon \mathbb{P}(\boldsymbol{X} \in B(\boldsymbol{x}, r)) > 0 \quad \textit{for all } r > 0\}.$$

Then $\mathcal{X}_0$ is the smallest closed set with $\mathbb{P}(\boldsymbol{X} \in \mathcal{X}_0^c) = 0$ and

$$f_\theta(\boldsymbol{x}) = \lim_{r \downarrow 0} \frac{P_\theta(B(\boldsymbol{x}, r))}{M(B(\boldsymbol{x}, r))} \quad \textit{for all } \theta \in \mathcal{Y}, \boldsymbol{x} \in \mathcal{X}_0. \tag{1.5}$$

To estimate P_θ, we use the *empirical measure* of the points $\boldsymbol{X}_i, i \in \mathcal{G}_\theta$,

$$\widehat{P}_\theta(B) := \frac{\#\{i \in \mathcal{G}_\theta \colon \boldsymbol{X}_i \in B\}}{N_\theta} \qquad \text{for } B \subset \mathcal{X}.$$

Now define for fixed integer $k \le n$ and any $\boldsymbol{x} \in \mathcal{X}$

$$\widehat{r}_{k,n}(\boldsymbol{x}) = \widehat{r}_{k,n}(\boldsymbol{x}, \mathcal{D}) := \min\big\{r \ge 0 \colon \#\{i \le n \colon \boldsymbol{X}_i \in B(\boldsymbol{x}, r)\} \ge k\big\}$$

such that $B(\boldsymbol{x}, \widehat{r}_{k,n}(\boldsymbol{x}))$ is the smallest ball centered at $\boldsymbol{x}$, which covers at least k training vectors $\boldsymbol{X}_i$. These are the k *nearest neighbors* of $\boldsymbol{x}$. Next we define

$$\widehat{f}_\theta(\boldsymbol{x}) := \frac{\widehat{P}_\theta\big(B(\boldsymbol{x}, \widehat{r}_{k,n}(\boldsymbol{x}))\big)}{M\big(B(\boldsymbol{x}, \widehat{r}_{k,n}(\boldsymbol{x}))\big)}.$$

Since $M\big(B(\boldsymbol{x}, \widehat{r}_{k,n}(\boldsymbol{x}))\big)$ is the same for all classes $\theta \in \mathcal{Y}$, we end up with an estimator of the form

$$\widehat{Y}_k(\boldsymbol{x}) \in \operatorname*{arg\,max}_{\theta \in \mathcal{Y}} \widehat{w}_\theta \widehat{P}_\theta\big(B(\boldsymbol{x}, \widehat{r}_{k,n}(\boldsymbol{x}))\big).$$

For $\widehat{w}_\theta = N_\theta / n$, this can be written as

$$\widehat{Y}_k(\boldsymbol{x}) \in \operatorname*{arg\,max}_{\theta \in \mathcal{Y}} \#\big\{i \in \mathcal{G}_\theta \colon \boldsymbol{X}_i \in B(\boldsymbol{x}, \widehat{r}_{k,n}(\boldsymbol{x}))\big\}.$$

This means we use majority vote among the k nearest neighbors to classify $\boldsymbol{X}$.

Proof of Lemma 1.2. First we show that $\mathcal{X}_0$ is closed. Let $\boldsymbol{x} \in \mathcal{X} \setminus \mathcal{X}_0$. Then there is a $r > 0$ such that $\mathbb{P}(\boldsymbol{X} \in U(\boldsymbol{x}, r)) = 0$. Now let $\boldsymbol{y} \in U(\boldsymbol{x}, r)$. Then $U(\boldsymbol{y}, \tilde{r}) \subset U(\boldsymbol{x}, r)$ with $\tilde{r} = r - d(\boldsymbol{x}, \boldsymbol{y}) > 0$. But this implies that $\mathbb{P}(\boldsymbol{X} \in U(\boldsymbol{y}, \tilde{r})) = 0$, and thus $\boldsymbol{y} \notin \mathcal{X}_0$. Therefore $U(\boldsymbol{x}, r) \cap \mathcal{X}_0 = \emptyset$. Since the choice of $\boldsymbol{x} \in \mathcal{X} \setminus \mathcal{X}_0$ was arbitrary, this implies that $\mathcal{X}_0$ is closed.

Let $\mathcal{X}_*$ be a countable and dense subset of $\mathcal{X}$. For each $\boldsymbol{x} \in \mathcal{X}_0^c$ there exists an $r > 0$ such that $\mathbb{P}(\boldsymbol{X} \in B(\boldsymbol{x}, r)) = 0$. Then there is a $\boldsymbol{x}_* \in \mathcal{X}_*$ such that $d(\boldsymbol{x}, \boldsymbol{x}_*) < r/2$, so $\boldsymbol{x} \in B(\boldsymbol{x}_*, r_*) \subset B(\boldsymbol{x}, r)$ for any $r_* \in (r/2, r - d(\boldsymbol{x}, \boldsymbol{x}_*)) \cap \mathbb{Q}$. This construction shows that

$$\mathcal{X}_0^c \subset \bigcup \big\{B(\boldsymbol{x}_*, r_*) \colon \boldsymbol{x}_* \in \mathcal{X}_*, 0 < r_* \in \mathbb{Q}, \mathbb{P}\big(\boldsymbol{X} \in B(\boldsymbol{X}_*, r_*)\big) = 0\big\}.$$

Since the latter union is countable, it has measure zero as well.

Now suppose that there is a closed set $\mathcal{X}_0' \subsetneq \mathcal{X}_0$ satisfying $\mathbb{P}(\boldsymbol{X} \in \mathcal{X}_0'^c) = 0$. Then for any $\boldsymbol{x} \in \mathcal{X}_0 \setminus \mathcal{X}_0'$ there is a $r > 0$ such that $B(\boldsymbol{x}, r) \cap \mathcal{X}_0' = \emptyset$ and so $\mathbb{P}(\boldsymbol{X} \in B(\boldsymbol{x}, r)) = 0$, which contradicts the definition of $\mathcal{X}_0$. Thus $\mathcal{X}_0$ is the smallest closed set with $\mathbb{P}(\boldsymbol{X} \in \mathcal{X}_0^c) = 0$.

Next we show that (1.5) holds. Since

$$P_\theta\big(B(\boldsymbol{x}, r)\big) = \int\limits_{B(\boldsymbol{x}, r)} f_\theta(\boldsymbol{y})\, M(\mathrm{d}\boldsymbol{y}) \begin{cases} \leq \sup\limits_{\boldsymbol{y} \in B(\boldsymbol{x}, r)} f_\theta(\boldsymbol{y}) M\big(B(\boldsymbol{x}, r)\big) \\ \geq \inf\limits_{\boldsymbol{y} \in B(\boldsymbol{x}, r)} f_\theta(\boldsymbol{y}) M\big(B(\boldsymbol{x}, r)\big), \end{cases}$$

it follows from the continuity of f that

$$\lim_{r \downarrow 0} \frac{P_\theta\big(B(\boldsymbol{x}, r)\big)}{M\big(B(\boldsymbol{x}, r)\big)} = f_\theta(\boldsymbol{x}).$$

□

Weighted Nearest Neighbors

In the previous paragraph the k-nearest neighbors of the observation $\boldsymbol{X}$ were all weighted equally. However, it would be reasonable to assign larger weights to training observations which are closer to $\boldsymbol{X}$. Now we first order the training data according to their distance to $\boldsymbol{X}$ and then we assign descending weights to them. Let

$$W_n(1) \geq W_n(2) \geq \ldots \geq W_n(n) \geq 0$$

be weights and

$$\widehat{R}(\boldsymbol{x}, \boldsymbol{X}_i) := \#\big\{j \leq n \colon d(\boldsymbol{x}, \boldsymbol{X}_j) \leq d(\boldsymbol{x}, \boldsymbol{X}_i)\big\}$$

the rank of training observation $\boldsymbol{X}_i$ according to its distance to $\boldsymbol{x}$. If two or more training observations are at equal distance to $\boldsymbol{x}$ we could average the weights or use randomization. However, we assume that

$$M\{\boldsymbol{y} \in \mathcal{X} \colon d(\boldsymbol{x}, \boldsymbol{y}) = r\} = 0 \quad \text{for all } \boldsymbol{x} \in \mathcal{X}, r \geq 0 \tag{1.6}$$

to avoid this problem. We define the *weighted nearest neighbor classifier* by

$$\widehat{Y}_{wnn}(\boldsymbol{x}) \in \operatorname*{arg\,max}_{\theta \in \mathcal{Y}} \sum_{i \in \mathcal{G}_\theta} W_n\big(\widehat{R}(\boldsymbol{x}, \boldsymbol{X}_i)\big).$$

1.1.3. Estimation of Missclassification Rates

To judge a certain classifier one could estimate the missclassification rates

$$R_\theta = \mathbb{P}\big(\widehat{Y}(\boldsymbol{X}, \mathcal{D}) \neq Y \mid Y = \theta, \mathcal{D}\big)$$

using cross-validation, i.e. with the estimator

$$\widehat{R}_\theta = \frac{\#\{i \in \mathcal{G}_\theta \colon \widehat{Y}(\boldsymbol{X}_i, \mathcal{D}_i) \neq Y_i\}}{N_\theta},$$

where $\mathcal{D}_i$ denotes the training data without observation $(\boldsymbol{X}_i, Y_i)$, and compare them with the missclassification rates of the optimal classifier.

1.2. From Classifiers to P-Values

A drawback of point estimators is the lack of information about confidence. To get such information we could use a Bayesian approach and calculate the *posterior distribution* of Y given $\boldsymbol{X}$, i.e. the *posterior weights*

$$w_\theta(\boldsymbol{X}) := \mathbb{P}(Y = \theta \mid \boldsymbol{X}).$$

By Lemma 1.1, a classifier $\widehat{Y}^*$ satisfying

$$\widehat{Y}^*(\boldsymbol{x}) \in \operatorname*{arg\,max}_{\theta \in \mathcal{Y}} w_\theta(\boldsymbol{x})$$

is optimal in the sense, that it minimizes the risk $R(\widehat{Y}) = \mathbb{P}(\widehat{Y}(\boldsymbol{X}) \neq Y)$. Thus we can now compute the conditional risk

$$\mathbb{P}(\widehat{Y}^*(\boldsymbol{X}) \neq Y \mid \boldsymbol{X} = \boldsymbol{x}) = 1 - \max_{\theta \in \mathcal{Y}} w_\theta(\boldsymbol{x}),$$

which gives us information about confidence of $\widehat{Y}^*$.

However a drawback of the posterior probabilities is, that the posterior weights $w_\theta(\boldsymbol{X})$ depend sensitively on the prior weights w_θ, i.e. small changes in w_θ can lead to totally different $w_\theta(\boldsymbol{X})$, which we illustrate in Example 1.1. In addition, classes with small prior weights w_θ tend to be ignored by the classifier $\widehat{Y}^*$, and so the class-dependent risk $\mathbb{P}(\widehat{Y}^*(\boldsymbol{X}) \neq Y \mid Y = \theta)$ may be rather large for some classes θ. Moreover, in some studies the class labels are not random, but predetermined by the study design. For example in a case-control study, one recruits a certain number of diseased individuals and a certain number of healthy individuals. Furthermore, in medical studies the prior probabilities can change over time or differ geographically, while the distributions P_θ are reasonably assumed to be universal. Another problem arises, if the future observation $(\boldsymbol{X}, Y)$ belongs to a so far unknown class $\theta \notin \mathcal{Y}$.

In the daily routine one often uses a process of elimination to classify objects. In our context, this means that we exclude certain classes $\theta \in \mathcal{Y}$ and finally give a set of plausible candidates for Y. In other words we treat Y temporarily as an unknown fixed parameter and compute for each class $\theta \in \mathcal{Y}$ a p-value $\pi_\theta(\boldsymbol{X})$ or $\pi_\theta(\boldsymbol{X}, \mathcal{D})$ for the null hypothesis that $Y = \theta$. In the ideal case, where the joint distribution of $(\boldsymbol{X}, Y)$ is known, this means $\pi_\theta \colon \mathcal{X} \to [0, 1]$ satisfies

$$\mathbb{P}(\pi_\theta(\boldsymbol{X}) \leq \alpha \mid Y = \theta) \leq \alpha \quad \text{for all } \alpha \in (0, 1). \tag{1.7}$$

Given such p-values π_θ, the set

$$\widehat{\mathcal{Y}}_\alpha(\boldsymbol{X}) := \{\theta \in \mathcal{Y} \colon \pi_\theta(\boldsymbol{X}) > \alpha\}$$

is a $(1 - \alpha)$-prediction region for Y, i.e.

$$\mathbb{P}(Y \in \widehat{\mathcal{Y}}_\alpha(\boldsymbol{X}) \mid Y = \theta) \geq 1 - \alpha \quad \text{for any } \theta \in \mathcal{Y}, \alpha \in (0, 1).$$

Thus we can exclude the classes $\theta \notin \widehat{\mathcal{Y}}_\alpha(\boldsymbol{X})$ with confidence $1 - \alpha$. If there is only one $\theta \in \widehat{\mathcal{Y}}_\alpha(\boldsymbol{X})$, we have classified $\boldsymbol{X}$ uniquely with confidence $1 - \alpha$.

Since we compute p-values for multiple null hypotheses, one could expect that we get a multiple testing problem. However, the observation $\boldsymbol{X}$ belongs to only one class and therefore at most one null hypothesis is true.

In the realistic case, where the joint distribution of $(\boldsymbol{X}, Y)$ is unknown, we compute p-values $\pi_\theta(\boldsymbol{X}, \mathcal{D})$ depending on the current feature vector $\boldsymbol{X}$ as well as on the training data $\mathcal{D}$. In this case, condition (1.7) can be extended in two ways:

$$\mathbb{P}(\pi_\theta(\boldsymbol{X}, \mathcal{D}) \le \alpha \mid Y = \theta) \le \alpha, \tag{1.8}$$

$$\mathbb{P}(\pi_\theta(\boldsymbol{X}, \mathcal{D}) \le \alpha \mid Y = \theta, \mathcal{D}) \le \alpha + \mathrm{o}_{\mathrm{p}}(1) \quad \text{as } n \to \infty, \tag{1.9}$$

for any $\theta \in \mathcal{Y}$ and $\alpha \in (0, 1)$. Condition (1.8) corresponds to "single use" and Condition (1.9) to "multiple use". Suppose that we construct p-values $\pi_\theta(\cdot, \mathcal{D})$ based on one training data set $\mathcal{D}$ and classify many future observations $(\tilde{\boldsymbol{X}}, \tilde{Y})$. Then the relative number of future observations with $\tilde{Y} = b$ and $\pi_\theta(\tilde{\boldsymbol{X}}, \mathcal{D}) \le \alpha$ is close to

$$w_b \mathbb{P}(\pi_\theta(\boldsymbol{X}, \mathcal{D}) \le \alpha \mid Y = b, \mathcal{D}),$$

a random quantity depending on the training data $\mathcal{D}$.

Example 1.1. For the following one-dimensional example let $L = 2$, $P_1 = \text{Gamma}(3, 1)$ and $P_2 = \text{Gamma}(6, 1)$. Figure 1.1 illustrates how the optimal point predictor $Y^*(x)$ and the posterior weights $w_\theta(x)$ depend on the prior probabilities w_θ. It shows $w_2(x)$ for $w_2/w_1 = 10, 1.5, 1, 0.67, 0.1$ (from left to right). The corresponding boundaries of $Y^*(x)$ are drawn as vertical lines.

Alternatively, we could calculate p-values which do not depend on the prior probabilities w_θ. Since $w_2(x)$ is increasing in x, we define the p-values

$$\pi_1(x) := \mathbb{P}(X \ge x \mid Y = 1),$$
$$\pi_2(x) := \mathbb{P}(X \le x \mid Y = 2).$$

If $\pi_\theta(X) \le \alpha$ we claim with confidence $1 - \alpha$ that $Y \ne \theta$. Figure 1.2 shows the p-value functions $\pi_1(x)$ and $\pi_2(x)$. In addition, the three regions where $\widehat{\mathcal{Y}}_{0.1}(x) = \{1\}, \{2\}, \{1, 2\}$ are marked.

1.3. Optimal P-Values as Benchmark

In this section we suppose that the distributions $P_1, P_2, \ldots, P_L$ have known densities $f_1, f_2, \ldots, f_L > 0$ with respect to some measure M on $\mathcal{X}$. Then the marginal distribution of $\boldsymbol{X}$ has density

$$f(\boldsymbol{x}) := \sum_{\theta \in \mathcal{Y}} w_\theta f_\theta(\boldsymbol{x})$$

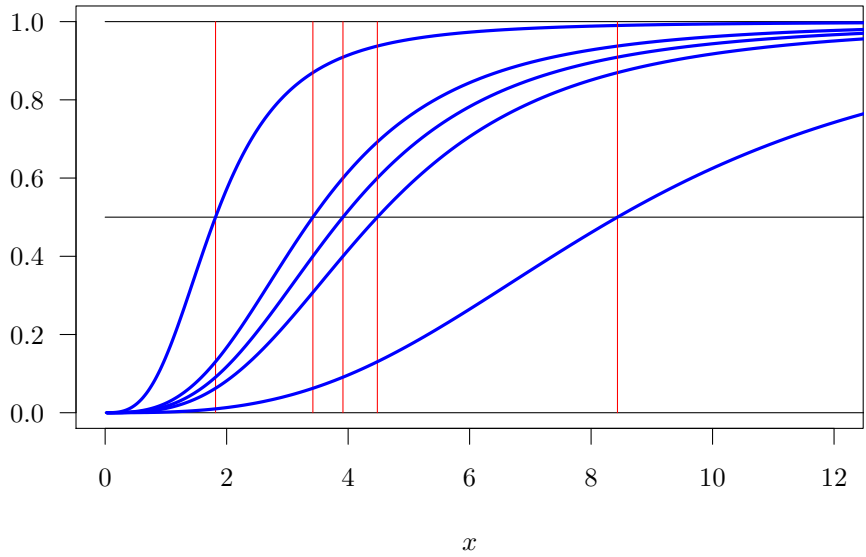

Figure 1.1.: Posterior weights $w_2(x)$ for different ratios of prior probabilities w_2/w_1.

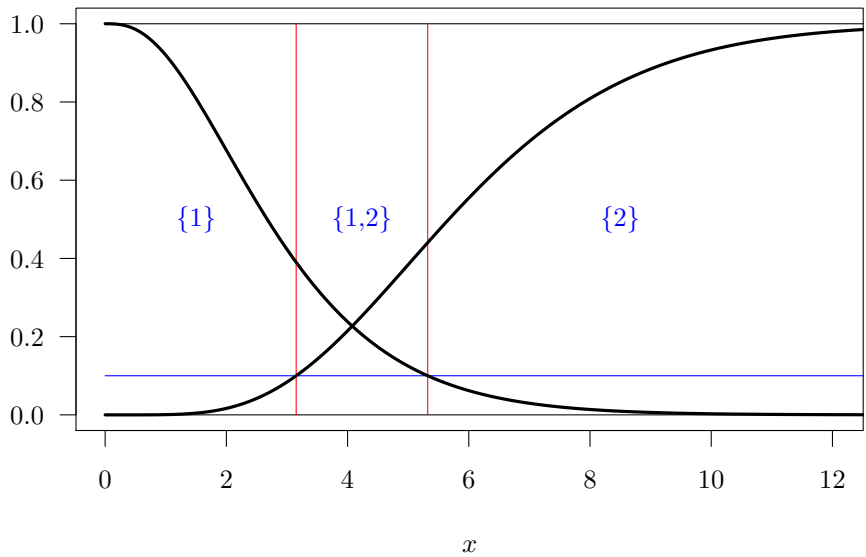

Figure 1.2.: P-value functions for class memberships.

with respect to M and

$$w_\theta(\boldsymbol{x}) = \frac{w_\theta f_\theta(\boldsymbol{x})}{f(\boldsymbol{x})}.$$

Let $\boldsymbol{\pi} = (\pi_\theta)_{\theta\in\mathcal{Y}}$ consist of p-values π_θ satisfying (1.8). Given the latter constraint we want to provide small p-values and small prediction regions. Therefore we use the following measures of risk:

$$\mathcal{R}(\boldsymbol{\pi}) := \mathbb{E}\big(\sum_{\theta\in\mathcal{Y}} \pi_\theta(\boldsymbol{X})\big),$$
$$\mathcal{R}_\alpha(\boldsymbol{\pi}) := \mathbb{E}\big(\#\widehat{\mathcal{Y}}_\alpha(\boldsymbol{X})\big).$$

Lemma 1.3. *Define* $\mathcal{R}_\alpha(\pi_\theta) := \mathbb{P}(\pi_\theta(\boldsymbol{X}) > \alpha)$. *Then*

$$\mathcal{R}(\boldsymbol{\pi}) = \int_0^1 \mathcal{R}_\alpha(\boldsymbol{\pi})\,\mathrm{d}\alpha$$

and

$$\mathcal{R}_\alpha(\boldsymbol{\pi}) = \sum_{\theta\in\mathcal{Y}} \mathcal{R}_\alpha(\pi_\theta).$$

Proof. Note that

$$\sum_{\theta\in\mathcal{Y}} \mathcal{R}_\alpha(\pi_\theta) = \mathbb{E}\big(\#\{\theta\in\mathcal{Y}\colon \pi_\theta(\boldsymbol{X}) > \alpha\}\big) = \mathcal{R}_\alpha(\boldsymbol{\pi})$$

and

$$\mathcal{R}(\boldsymbol{\pi}) = \sum_{\theta\in\mathcal{Y}} \int_0^1 \mathbb{P}\big(\pi_\theta(\boldsymbol{X}) > \alpha\big)\,\mathrm{d}\alpha = \int_0^1 \mathcal{R}_\alpha(\boldsymbol{\pi})\,\mathrm{d}\alpha.$$

□

In view of the preceding lemma, we focus on minimizing $\mathcal{R}_\alpha(\pi_\theta)$ for arbitrary fixed $\theta\in\mathcal{Y}$ and $\alpha\in(0,1)$ under the constraint (1.8).

Lemma 1.4. *Let* $\mathcal{L}((f_\theta/f)(\boldsymbol{X}))$ *be continuous. Then the p-value*

$$\pi_\theta^*(\boldsymbol{x}) := P_\theta\big\{\boldsymbol{z}\in\mathcal{X}\colon (f_\theta/f)(\boldsymbol{z}) \le (f_\theta/f)(\boldsymbol{x})\big\}$$

is optimal, in the sense that $\mathcal{R}_\alpha(\pi_\theta^*)$ *is minimal for each* $\alpha\in(0,1)$.

Proof. We consider

$$\varphi(\boldsymbol{x}) := \mathbb{1}\{\pi_\theta(\boldsymbol{x}) \le \alpha\}$$

as a level-α test of the null-hypothesis P_θ versus the alternative hypothesis $P = \sum_{b\in\mathcal{Y}} w_b P_b$ and maximize the power

$$\mathbb{E}\varphi(\boldsymbol{X}) = \int \varphi(\boldsymbol{x})\, P(\mathrm{d}\boldsymbol{x})$$

subject to the condition

$$\mathbb{E}_\theta\varphi(\boldsymbol{X}) := \int \varphi(\boldsymbol{x})\, P_\theta(\mathrm{d}\boldsymbol{x}) \le \alpha.$$

The Neyman-Pearson Lemma (Theorem A.2) yields that $\mathbb{E}\varphi(\boldsymbol{X})$ is maximal for

$$\varphi(\boldsymbol{x}) = \begin{cases} 1 & \text{if } (f/f_\theta)(\boldsymbol{x}) > c_\theta, \\ \gamma_\theta & \text{if } (f/f_\theta)(\boldsymbol{x}) = c_\theta, \\ 0 & \text{if } (f/f_\theta)(\boldsymbol{x}) < c_\theta \end{cases}$$

with $c_\theta \in [0,\infty]$ and $\gamma_\theta \in [0,1]$ such that

$$\begin{aligned}\mathbb{E}_\theta\varphi(\boldsymbol{X}) &= P_\theta\big\{\boldsymbol{x}\in\mathcal{X}\colon (f/f_\theta)(\boldsymbol{x}) > c_\theta\big\} + \gamma_\theta P_\theta\big\{\boldsymbol{x}\in\mathcal{X}\colon (f/f_\theta)(\boldsymbol{x}) = c_\theta\big\} \\ &= \alpha.\end{aligned}$$

Since $\mathcal{L}\big((f_\theta/f)(\boldsymbol{X})\big)$ is continuous, γ_θ can be chosen arbitrarily. With $\gamma_\theta = 1$ and

$$c_\theta := \min\Big\{c\colon P_\theta\big\{\boldsymbol{x}\in\mathcal{X}\colon (f/f_\theta)(\boldsymbol{x}) \ge c\big\} \le \alpha\Big\}$$

we get

$$\begin{aligned}\varphi(\boldsymbol{x}) &= \mathbb{1}\big\{(f/f_\theta)(\boldsymbol{x}) \ge c_\theta\big\} \\ &= \mathbb{1}\big\{\pi_\theta^*(x) \le \alpha\big\}.\end{aligned}$$

Thus $\varphi(\boldsymbol{x}) = \mathbb{1}\{\pi_\theta^*(\boldsymbol{x}) \le \alpha\}$ maximizes $\mathbb{E}\varphi(\boldsymbol{x})$ and minimizes $\mathcal{R}_\alpha(\pi_\theta) = 1 - \mathbb{E}\varphi(\boldsymbol{x})$. □

Two other representations of $\pi_\theta^*(\boldsymbol{x})$ are given by

$$\begin{aligned}\pi_\theta^*(\boldsymbol{x}) &= P_\theta\{\boldsymbol{z}\in\mathcal{X}\colon w_\theta(\boldsymbol{z}) \le w_\theta(\boldsymbol{x})\} \\ &= P_\theta\{\boldsymbol{z}\in\mathcal{X}\colon T_\theta^*(\boldsymbol{z}) \ge T_\theta^*(\boldsymbol{x})\}\end{aligned}$$

with

$$T_\theta^*(\boldsymbol{x}) := \sum_{b\ne\theta} \frac{w_{b,\theta} f_b(\boldsymbol{x})}{f_\theta(\boldsymbol{x})} \quad \text{and} \quad w_{b,\theta} := \Big(\sum_{c\ne\theta} w_c/w_b\Big)^{-1}.$$

Note that

$$\begin{aligned} T_\theta^*(\cdot) &= \frac{f(\cdot) - w_\theta f_\theta(\cdot)}{f_\theta(\cdot)(1 - w_\theta)} \\ &= \frac{1 - w_\theta(\cdot)}{\frac{f_\theta}{f}(\cdot)(1 - w_\theta)} \\ &= \frac{w_\theta(\cdot)^{-1} - 1}{w_\theta^{-1} - 1} \end{aligned}$$

is a negative monotonic transformation of $w_\theta(\cdot)$. The first representation shows that $\pi_\theta^*(\boldsymbol{x})$ is a non-decreasing function of $w_\theta(\boldsymbol{x})$. The second representation shows that the prior weight w_θ itself is irrelevant for the optimal p-value π_θ^*. Only the ratios w_c/w_b with $b, c \neq \theta$ matter. In particular, in case of $L = 2$ classes, the optimal p-values do not depend on the prior distribution of Y at all.

Example 1.2 (Standard model). Let $P_\theta = \mathcal{N}_d(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma})$ with a common covariance matrix $\boldsymbol{\Sigma}$. Then

$$T_\theta^*(\boldsymbol{x}) := \sum_{b \neq \theta} w_{b,\theta} \exp\left((\boldsymbol{x} - \boldsymbol{\mu}_{\theta,b})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_b - \boldsymbol{\mu}_\theta)\right) \tag{1.10}$$

with $\boldsymbol{\mu}_{\theta,b} = (\boldsymbol{\mu}_\theta + \boldsymbol{\mu}_b)/2$.

1.4. P-Values via Permutation Tests

Now we suppose that the joint distribution of $(\boldsymbol{X}, Y)$ is unknown and compute p-values $\pi_\theta(\boldsymbol{X}, \mathcal{D})$ and prediction regions

$$\widehat{\mathcal{Y}}_\alpha(\boldsymbol{X}, \mathcal{D}) := \{\theta \in \mathcal{Y} \colon \pi_\theta(\boldsymbol{X}, \mathcal{D}) > \alpha\}$$

depending on training data $\mathcal{D}$. We introduce nonparametric p-values for classification as they are given in Dümbgen et al. (2008).

For the remaining part of this thesis we consider the class labels $Y_1, Y_2, \ldots, Y_n$ as fixed while $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ and $(\boldsymbol{X}, Y)$ are independent with $\mathcal{L}(\boldsymbol{X}_i) = P_{Y_i}$. That way we can handle situations with stratified training data as well as the case of i.i.d. training data (via conditioning).

Further we assume that the distributions $P_1, P_2, \ldots, P_L$ have densities $f_1, f_2, \ldots, f_L > 0$ with respect to some measure M on $\mathcal{X}$ and that all group sizes N_θ are strictly positive. Asymptotic statements as in (1.9) are meant as

$$n \to \infty \quad \text{and} \quad N_b/n \to w_b \quad \text{for all } b \in \mathcal{Y}. \tag{1.11}$$

Let $I(1) < I(2) < \cdots < I(N_\theta)$ be the elements of $\mathcal{G}_\theta$ for a fixed class $\theta \in \mathcal{Y}$. Then $(\boldsymbol{X}, \boldsymbol{X}_{I(1)}, \boldsymbol{X}_{I(2)}, \ldots, \boldsymbol{X}_{I(N_\theta)})$ is exchangeable conditional on $Y = \theta$.

Thus we consider a test statistic $T_\theta(\boldsymbol{X}, \mathcal{D})$ which is symmetric in $(\boldsymbol{X}_{I(j)})_{j=1}^{N_\theta}$. We define $\mathcal{D}_i(\boldsymbol{x})$ to be the training data with $\boldsymbol{x}$ in place of $\boldsymbol{X}_i$. Then the nonparametric p-value

$$\pi_\theta(\boldsymbol{X}, \mathcal{D}) := \frac{\#\big\{i \in \mathcal{G}_\theta \colon T_\theta\big(\boldsymbol{X}_i, \mathcal{D}_i(\boldsymbol{X})\big) \geq T_\theta(\boldsymbol{X}, \mathcal{D})\big\} + 1}{N_\theta + 1} \tag{1.12}$$

satisfies (1.8). By definition, $\pi_\theta(\boldsymbol{X}, \mathcal{D}) \geq (N_\theta + 1)^{-1}$. Therefore this procedure is only useful, if $N_\theta + 1 \geq \alpha^{-1}$. For instance if $\alpha = 0.05$, N_θ should be at least 19.

Plug-In Statistic for Standard Model

In the standard model with $P_\theta = \mathcal{N}_d(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma})$ we replace the unknown mean vectors $\boldsymbol{\mu}_\theta$ and covariance matrix $\boldsymbol{\Sigma}$ in (1.10) with corresponding estimators. Note that the resulting p-values always satisfy (1.8), even if the underlying distributions P_θ are not Gaussian with common covariance matrix.

To compute $\pi_\theta(\boldsymbol{X}, \mathcal{D})$, we add the new observation $\boldsymbol{X}$ temporarily to class θ. But $\boldsymbol{X}$ may be an outlier with respect to the distribution P_θ. Therefore it is reasonable to use one of the robust M-estimators mentioned in Section 1.1.2.

Nearest Neighbors

Now we use the test statistic $T_\theta(\boldsymbol{x}, \mathcal{D}) = -w_\theta(\boldsymbol{x})$ and estimate $w_\theta(\boldsymbol{x})$ via nearest neighbors (cf. Section 1.1.2). For the k-nearest neighbors we get

$$\widehat{w}_\theta(\boldsymbol{x}, \mathcal{D}) := \frac{\widehat{w}_\theta \widehat{P}_\theta\big(B(\boldsymbol{x}, \widehat{r}_{k,n}(\boldsymbol{x}))\big)}{\sum\limits_{b \in \mathcal{Y}} \widehat{w}_b \widehat{P}_b\big(B(\boldsymbol{x}, \widehat{r}_{k,n}(\boldsymbol{x}))\big)}$$

with certain estimators $\widehat{w}_b = \widehat{w}_b(\mathcal{D})$ of w_b. In case of $\widehat{w}_b = N_b/n$, we get

$$\begin{aligned}\widehat{w}_\theta(\boldsymbol{x}, \mathcal{D}) &:= \frac{\#\{i \in \mathcal{G}_\theta \colon d(\boldsymbol{x}, \boldsymbol{X}_i) \leq \widehat{r}_{k,n}(\boldsymbol{x})\}}{\#\{i \leq n \colon d(\boldsymbol{x}, \boldsymbol{X}_i) \leq \widehat{r}_{k,n}(\boldsymbol{x})\}} \\ &= \frac{\sum\limits_{i=1}^{n} \mathbb{1}\{d(\boldsymbol{x}, \boldsymbol{X}_i) \leq \widehat{r}_{k,n}(\boldsymbol{x})\}\mathbb{1}\{Y_i = \theta\}}{\sum\limits_{i=1}^{n} \mathbb{1}\{d(\boldsymbol{x}, \boldsymbol{X}_i) \leq \widehat{r}_{k,n}(\boldsymbol{x})\}}.\end{aligned}$$

For the weighted nearest neighbors we get

$$\widehat{w}_\theta(\boldsymbol{x}, \mathcal{D}) := \frac{\sum\limits_{i=1}^{n} W_n(\widehat{R}(\boldsymbol{x}, \boldsymbol{X}_i))\mathbb{1}\{Y_i = \theta\}}{\sum\limits_{i=1}^{n} W_n(\widehat{R}(\boldsymbol{x}, \boldsymbol{X}_i))}.$$

Penalized Multicategory Logistic Regression

Let $\mathcal{X} = \mathbb{R}^d$ and $\boldsymbol{X}$ contain the values of d numerical or binary variables. We assume that

$$\mathbb{P}(Y = \theta \,|\, \boldsymbol{X} = \boldsymbol{x}) = \exp(a_\theta + \boldsymbol{b}_\theta^\top \boldsymbol{x}) \Big/ \sum_{z=1}^{L} \exp(a_z + \boldsymbol{b}_z^\top \boldsymbol{x})$$

for unknown parameters $a_z \in \mathbb{R}$ and $\boldsymbol{b}_z \in \mathbb{R}^d$, which we estimate with penalized maximum likelihood estimators $\widehat{a}_z(\mathcal{D})$ and $\widehat{\boldsymbol{b}}_z(\mathcal{D})$. To compute the p-values, we use the test statistic

$$T_\theta(\boldsymbol{x}, \mathcal{D}) = -\exp(\widehat{a}_\theta + \widehat{\boldsymbol{b}}_\theta^\top \boldsymbol{x}) \Big/ \sum_{z=1}^{L} \exp(\widehat{a}_z + \widehat{\boldsymbol{b}}_z^\top \boldsymbol{x}).$$

Technical details for the penalized multicategory logistic regression are given in Section 1.8.

1.5. Estimation of Separability

To estimate the separability of different classes by means of given p-values $\pi_\theta(\cdot, \cdot)$ we compute *cross-validated* p-values

$$\pi_\theta(\boldsymbol{X}_i, \mathcal{D}_i)$$

for $i = 1, 2, \ldots, n$ with $\mathcal{D}_i$ denoting the training data without observation $(\boldsymbol{X}_i, Y_i)$. We treat each training observation $(\boldsymbol{X}_i, Y_i)$ temporarily as a 'future' observation, which has to be classified with the remaining data $\mathcal{D}_i$. Then we could display these p-values graphically or compute the empirical conditional inclusion probabilities

$$\widehat{\mathcal{I}}_\alpha(b, \theta) := \frac{\#\{i \in \mathcal{G}_b \colon \theta \in \widehat{\mathcal{Y}}_\alpha(\boldsymbol{X}_i, \mathcal{D}_i)\}}{N_b}$$

and the empirical pattern probabilities

$$\widehat{\mathcal{P}}_\alpha(b, S) := \frac{\#\{i \in \mathcal{G}_b \colon \widehat{\mathcal{Y}}_\alpha(\boldsymbol{X}_i, \mathcal{D}_i) = S\}}{N_b}$$

for $b, \theta \in \mathcal{Y}$ and $S \subset \mathcal{Y}$. These numbers $\widehat{\mathcal{I}}_\alpha(b, \theta)$ and $\widehat{\mathcal{P}}_\alpha(b, S)$ can be interpreted as estimators of the conditional inclusion probabilities

$$\mathcal{I}_\alpha(b, \theta \mid \mathcal{D}) := \mathbb{P}(\theta \in \widehat{\mathcal{Y}}_\alpha(\boldsymbol{X}, \mathcal{D}) \mid Y = b, \mathcal{D})$$

and the conditional pattern probabilities

$$\mathcal{P}_\alpha(b, S \mid \mathcal{D}) := \mathbb{P}(\widehat{\mathcal{Y}}_\alpha(\boldsymbol{X}, \mathcal{D}) = S \mid Y = b, \mathcal{D}),$$

respectively.

To visualize the separability we plot for large group sizes N_b the empirical ROC curves

$$(0,1) \ni \alpha \mapsto 1 - \widehat{\mathcal{I}}_\alpha(b, \theta).$$

1.6. Asymptotic Properties

In this section we review the asymptotic results of Dümbgen et al. (2008) and Zumbrunnen (2009). For the plug-in statistic of the standard model we derive a central limit theorem in Section 3.4.

Throughout this section, asymptotic statements are to be understood within setting (1.11).

The following theorem implies that $\pi_\theta(\boldsymbol{X}, \mathcal{D})$ satisfies (1.9) under certain conditions on the underlying test statistic $T_\theta(\boldsymbol{X}, \mathcal{D})$. Furthermore it shows that the empirical conditional inclusion probabilities $\widehat{\mathcal{I}}_\alpha(b, \theta)$ and the empirical pattern probabilities $\widehat{\mathcal{P}}_\alpha(b, S)$ are consistent estimators of $\mathcal{I}_\alpha(b, \theta \mid \mathcal{D})$ and $\mathcal{P}_\alpha(b, S \mid \mathcal{D})$, respectively. Here and throughout $\to_p$ denotes convergence in probability.

Theorem 1.5. *Suppose that for fixed $\theta \in \mathcal{Y}$ there exists a test statistic T^o_θ on $\mathcal{X}$ satisfying the following two requirements:*

$$T_\theta(\boldsymbol{X}, \mathcal{D}) \to_{\mathrm{p}} T^o_\theta(\boldsymbol{X}), \tag{1.13}$$

$$\mathcal{L}\big(T^o_\theta(\boldsymbol{X})\big) \text{ is continuous.} \tag{1.14}$$

Then

$$\pi_\theta(\boldsymbol{X}, \mathcal{D}) \to_{\mathrm{p}} \pi^o_\theta(\boldsymbol{X}), \tag{1.15}$$

where

$$\pi^o_\theta(\boldsymbol{x}) := P_\theta\big\{\boldsymbol{z} \in \mathcal{X} \colon T^o_\theta(\boldsymbol{z}) \geq T^o_\theta(\boldsymbol{x})\big\}.$$

In particular, for arbitrary fixed $\alpha \in (0,1)$,

$$\mathcal{R}_\alpha\big(\pi_\theta(\cdot, \mathcal{D})\big) \to_{\mathrm{p}} \mathcal{R}_\alpha(\pi^o_\theta), \tag{1.16}$$

$$\left.\begin{matrix} \mathcal{I}_\alpha(b, \theta \mid \mathcal{D}) \\ \widehat{\mathcal{I}}_\alpha(b, \theta) \end{matrix}\right\} \to_{\mathrm{p}} \mathbb{P}(\pi^o_\theta(\boldsymbol{X}) > \alpha \mid Y = b) \quad \text{for each } b \in \mathcal{Y} \tag{1.17}$$

and

$$\left.\begin{array}{l}\mathcal{P}_\alpha(b, S \mid \mathcal{D}) \\ \widehat{\mathcal{P}}_\alpha(b, S)\end{array}\right\} \to_{\mathrm{p}} \mathbb{P}(\widehat{\mathcal{Y}}_\alpha^o(\boldsymbol{X}) = S) \quad \textit{for each } b \in \mathcal{Y} \textit{ and } S \subset \mathcal{Y}, \tag{1.18}$$

where $\widehat{\mathcal{Y}}_\alpha^o(\boldsymbol{X}) := \{\theta \in \mathcal{Y} \colon \pi_\theta^o(\boldsymbol{X}) > \alpha\}$.

The proof of this theorem can be found in Dümbgen et al. (2008), Theorem 3.1. The p-value $\pi_\theta(\cdot, \mathcal{D})$ is asymptotically optimal if T_θ^o is a strictly increasing transformation of T_θ^*. The following lemmata show that this is the case in different situations.

Plug-In Statistic for Standard Gaussian Model

Lemma 1.6. *Conditions* (1.13) *and* (1.14) *are satisfied with* $T_\theta^o = T_\theta^*$ *in case of the plug-in rule for the homoscedastic Gaussian model, provided that* $\mathbb{E}(\|\boldsymbol{X}\|^2) < \infty$ *and* $\mathcal{L}(\boldsymbol{X})$ *has a Lebesgue density.*

For the proof we refer to Dümbgen et al. (2008), Lemma 3.2.

Nearest Neighbors

Lemma 1.7. *Suppose that* $(\mathcal{X}, d)$ *is a separable metric space with a measure* M *satisfying* (1.6) *and that all densities* f_b, $b \in \mathcal{Y}$, *are continuous on* $\mathcal{X}$. *Then for the weighted nearest-neighbor rule with weights satisfying*

$$\sum_{i\colon i \geq \varepsilon n} W_n(i) \to 0 \quad \textit{for all } \varepsilon > 0, \tag{1.19}$$

$$\sum_{i=1}^{n} W_n(i) = 1 \quad \textit{for all } n \in \mathbb{N}, \tag{1.20}$$

$$W_n(1) \to 0, \tag{1.21}$$

the assumptions of Theorem 1.5 are satisfied with $T_\theta^o = T_\theta^*$.

The proof of this lemma can be found in Zumbrunnen (2009), Theorem 3.1. Note that the k-nearest neighbor rule with $\widehat{w}_\theta = N_\theta/n$ satisfies the conditions of the previous theorem, provided that

$$k = k(n) \to \infty \quad \text{and} \quad k/n \to 0.$$

Often different variables of a data set are measured on different scales. To take this into account, one could use the Mahalanobis distance, which is scale-invariant and data-driven.

Lemma 1.8. *Let $\mathcal{X}$ be an open subset of $\mathbb{R}^d$ and f_b, $b \in \mathcal{Y}$ continuous Lebesgue densities. Suppose that $\mathbb{E}\left(\|\boldsymbol{X}\|^2\right) < \infty$ and let $\widehat{\boldsymbol{\Sigma}}$ be a consistent estimator of the nonsingular matrix $\boldsymbol{\Sigma}_0 := \sum_{\theta\in\Theta} w_\theta \operatorname{Var}(\boldsymbol{X} \mid Y = \theta)$. Then in case of the weighted nearest-neighbor rule with the Mahalanobis distance $D_{\widehat{\boldsymbol{\Sigma}}}$ and weights satisfying* (1.19)–(1.21)*, the assumptions of Theorem 1.5 are satisfied with $T_\theta^o = T_\theta^*$.*

The proof of this lemma is given in Zumbrunnen (2009), Theorem 3.10.

1.7. Implementation in `pvclass`

The p-values for classification are implemented in the package `pvclass` (Zumbrunnen and Dümbgen, 2011). It was written in the `R` programming system (R Core Team, 2014) and depends on the recommended package `Matrix` (Bates and Maechler, 2010).

The main functions of `pvclass` compute p-values for the potential class memberships of new observations (`pvs`) as well as cross-validated p-values for training data (`cvpvs`). With the function `analyze.pvs`, the package `pvclass` also provides graphical displays and quantitative analyses of the p-values.

The test statistics of Section 1.4 are available in the package `pvclass`. It should be stressed however that users could easily implement test statistics corresponding to their own favorite classifier (e.g. neuronal nets).

To estimate the parameters we use N_θ/n for w_θ and the standard estimator for $\boldsymbol{\mu}_\theta$. For $\boldsymbol{\Sigma}$ the package `pvclass` offers the standard estimator as well as the more robust M-estimators $\widehat{\boldsymbol{\Sigma}}_M$ and $\widehat{\boldsymbol{\Sigma}}_{sym}$. The estimator $\widehat{\boldsymbol{\Sigma}}_{sym}$ requires that the observations $\boldsymbol{X}_i$ are pairwise different within groups. Otherwise, if an observation occurs more than once, `pvclass` uses only the first to calculate $\widehat{\boldsymbol{\Sigma}}_{sym}$.

For the nearest neighbor methods, `pvclass` offers besides the fixed Euclidean distance also two data-driven distances which are scale invariant. The Mahalanobis distance with respect to the estimated covariance matrix $\widehat{\boldsymbol{\Sigma}}$ as defined in Section 1.1.1 and the data driven Euclidean distance where we divide each component of $\boldsymbol{X}$ by its sample standard deviation and then use the Euclidean distance.

For the weighted nearest neighbors `pvclass` provides the linear weight function

$$W_n(i) = \max(1 - (i/n)/\tau, 0),$$

and the exponential weight function

$$W_n(i) = (1 - i/n)^\tau.$$

Alternatively one can specify the weights with an n dimensional vector W. For the exponential weight function τ should be in $(0, 1]$ and for the linear weight function it should be greater than 1.

Details for the test statistic based on penalized multicategory logistic regression are given in Section 1.8.

1.7.1. Shortcut

To reduce computation time, we add $(\boldsymbol{X}, \theta)$ to the training data before we judge the plausibility of the class label θ for a new observation $\boldsymbol{X}$, i.e. we replace $T_\theta(\boldsymbol{X}, \mathcal{D})$ and $T_\theta(\boldsymbol{X}_i, \mathcal{D}_i(\boldsymbol{X}))$ in Definition (1.12) by $T_\theta(\boldsymbol{X}, \mathcal{D}(\boldsymbol{X}, \theta))$ and $T_\theta(\boldsymbol{X}_i, \mathcal{D}(\boldsymbol{X}, \theta))$, respectively, where $\mathcal{D}(\boldsymbol{X}, \theta)$ denotes the training data extended by $(\boldsymbol{X}, \theta)$. Then we end up with the p-value

$$\pi_\theta(\boldsymbol{X}, \mathcal{D}(\boldsymbol{X}, \theta)) = \frac{\#\{i \in \mathcal{G}_\theta \colon T_\theta(\boldsymbol{X}_i, \mathcal{D}(\boldsymbol{X}, \theta)) \geq T_\theta(\boldsymbol{X}, \mathcal{D}(\boldsymbol{X}, \theta))\} + 1}{N_\theta + 1}.$$

To compute $\pi_\theta(\boldsymbol{X}, \mathcal{D}(\boldsymbol{X}, \theta))$ it suffices to evaluate $T_\theta(\cdot, \mathcal{D}(\boldsymbol{X}, \theta))$ at the $N_\theta + 1$ points $\boldsymbol{X}$ and $\boldsymbol{X}_i, i \in \mathcal{G}_\theta$. One can show that this p-value satisfies (1.8) and the conclusions of Theorems 1.5 and 3.11 remain true.

1.7.2. Data Example 'buerk'

To illustrate the main functions of `pvclass` we use the data set `buerk` provided by `pvclass`. It was collected by Prof. Dr. Conny Georg Bürk at the university hospital in Lübeck and contains data of 21'556 surgeries in a certain time period (end of the nineties). Besides the mortality and the morbidity it contains 21 variables describing the condition of the patient and the surgery. All collected variables can be found in Table 1.1.

We use the mortality as class label Y. The original data set contains 21'556 observations. To get a smaller data set, which is easier to handle, we take all 662 observations with $Y = 1$ and choose randomly $3 \cdot 662$ observations with $Y = 0$. For the test data set we choose 100 observations from each class. So we end up with a training data set containing 2448 observations, whereof 562 belong to class 1.

```
R> library(pvclass)
R> data(buerk)
R> set.seed(0)
R> X.raw <- as.matrix(buerk[, 1:21])
R> Y.raw <- buerk[, 22]
R> n0.raw <- sum(1 - Y.raw)
R> n1 <- sum(Y.raw)
R> n0 <- 3 * n1
```

Variable	Meaning
Y	Mortality (1 = deceased, 0 = survived)
X(1)	Age in years
X(2)	Sex (1 = female, 0 = male)
X(3)	ASA-Score (American Society of Anesthesiologists), describes the physical condition on an ordinal scale 1 = A normal healthy patient, 2 = A patient with mild systemic disease, 3 = A patient with severe systemic disease, 4 = A patient with severe systemic disease that is a constant threat to life, 5 = A moribund patient who is not expected to survive without the operation, 6 = A declared brain-dead patient whose organs are being removed for donor purposes
X(4)	Risk factor: cerebral (1 = yes, 0 = no)
X(5)	Risk factor: cardiovascular (1 = yes, 0 = no)
X(6)	Risk factor: pulmonary (1 = yes, 0 = no)
X(7)	Risk factor: renal (1 = yes, 0 = no)
X(8)	Risk factor: hepatic (1 = yes, 0 = no)
X(9)	Risk factor: immunological (1 = yes, 0 = no)
X(10)	Risk factor: metabolic (1 = yes, 0 = no)
X(11)	Risk factor: uncooperative, unreliable (1 = yes, 0 = no)
X(12)	Etiology: malignant (1 = yes, 0 = no)
X(13)	Etiology: vascular (1 = yes, 0 = no)
X(14)	Antibiotics therapy (1 = yes, 0 = no)
X(15)	Surgery indicated (1 = yes, 0 = no)
X(16)	Emergency operation (1 = yes, 0 = no)
X(17)	Surgery time in minutes
X(18)	Septic surgery (1 = yes, 0 = no)
X(19)	Experienced surgeon, i.e. senior physician (1 = yes, 0 = no)
X(20)	Blood transfusion necessary (1 = yes, 0 = no)
X(21)	Intensive care necessary (1 = yes, 0 = no)

Table 1.1.: Variables in `buerk` data set

```
R> X0 <- X.raw[Y.raw == 0, ]
R> X1 <- X.raw[Y.raw == 1, ]
R> tmpi0 <- sample(1:n0.raw, size = 3 * n1, replace = FALSE)
R> tmpi1 <- sample(1:n1, size = n1, replace = FALSE)
R> Xtrain <- rbind(X0[tmpi0[1:(n0 - 100)], ],
                   X1[1:(n1 - 100), ])
R> Ytrain <- c(rep(0, n0 - 100), rep(1, n1 - 100))
R> Xtest <- rbind(X0[tmpi0[(n0 - 99):n0], ],
                  X1[(n1 - 99):n1, ])
R> Ytest <- c(rep(0, 100), rep(1, 100))
```

1.7.3. Main Functions

Classify new observations

The function `pvs` computes nonparametric p-values for the potential class memberships of new observations. It returns a matrix `PV` containing the p-values. Precisely, for each new observation `NewX[i, ]` and each class `b` the number `PV[i,b]` is a p-value for the null hypothesis that $Y[i] = b$. With the option `method` or using directly one of the functions `pvs.method` one can choose a test statistic.

For the following example we use the weighted nearest neighbor statistic with an exponential weight function and `tau = 10`.

```
R> PV <- pvs(NewX = Xtest, X = Xtrain, Y = Ytrain,
             method = 'wnn', wtype = 'exponential', tau = 10)
R> head(PV)

             0          1
[1,] 0.1738209 0.45470693
[2,] 0.6173821 0.06216696
[3,] 0.1213567 0.58081705
[4,] 0.8473768 0.01776199
[5,] 0.4043455 0.15808171
[6,] 0.2517223 0.34280639
```

Next we illustrate the p-values graphically with the function `analyze.pvs` using the first ten observations of each class.

```
R> analyze.pvs(pv=PV[c(1:10,101:110),], alpha = 0.05)
```

For each p-value a rectangle with an area proportional to the p-value is drawn, see Figure 1.3. The rectangle is blue if the p-value is greater than `alpha` and red otherwise. If we specify the class labels of the test data as in the next example, then the data are sorted by class and the class labels are

Figure 1.3.: Illustration of the p-values without indicating the class labels of the test data

shown in the plot, see Figure 1.4. Additionally ROC curves are plotted by default. We suppress this here with the argument `roc = FALSE`. An example of the ROC curve plot can be found in the next section.

```
R> analyze.pvs(pv=PV[c(1:10,101:110), ],
+               Y = Ytest[c(1:10,101:110)], roc = FALSE)
```

Cross-validated p-values

The function `cvpvs` returns a matrix `PV` containing cross-validated nonparametric p-values for the potential class memberships of the training data. Precisely, for each feature vector `X[i,]` and each class `b` the number `PV[i,b]` is a p-value for the null hypothesis that $Y[i] = b$.

For the following example we use the logistic regression with penalty parameter `tau.o = 2`.

```
R> PV.cv <- cvpvs(X = Xtrain, Y = Ytrain,
                  method = 'logreg', tau.o = 2)
R> PV.cv[1:3,]
```

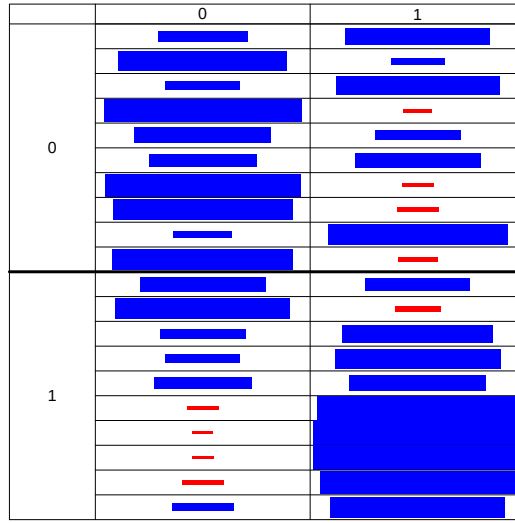

Figure 1.4.: Illustration of the p-values with class labels of the test data

```
          [,1]        [,2]
[1,] 0.9761400 0.001776199
[2,] 0.4172853 0.010657194
[3,] 0.4554613 0.010657194

R> PV.cv[2001:2003,]

             [,1]      [,2]
[1,] 0.002119767 0.7971530
[2,] 0.049284579 0.2740214
[3,] 0.010068892 0.6263345
```

The cross-validated p-values can be illustrated graphically the same way as the p-values for the new observations. If $L \leq 3$ the function `analyze.pvs` also prints the empirical pattern probabilities $\widehat{\mathcal{P}}_\alpha(b, S)$ for all subsets $S \subset \mathcal{Y}$. Otherwise it prints the empirical conditional inclusion probabilities $\widehat{\mathcal{I}}_\alpha(b, \theta)$ for all combinations of b and θ and the empirical pattern probabilities for $S = \emptyset, \mathcal{Y}$ and $\{\theta\}$ for all $\theta \in \mathcal{Y}$.

In the following example we suppress the plot of the p-values and get only the plot of the ROC curves, see Figure 1.5.

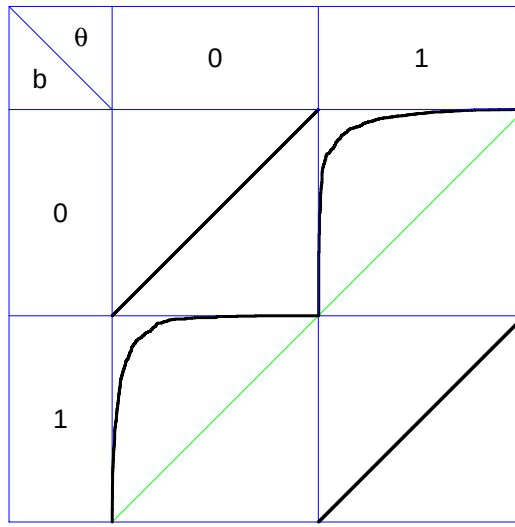

Figure 1.5.: ROC curves of the cross-validated p-values

```
R> analyze.pvs(pv = PV.cv, Y = Ytrain, pvplot = FALSE, cex=1.3)

b   P(b,{})   P(b,{1})   P(b,{2}) P(b,{1,2})
  1       0 0.78791092 0.04984093  0.1622481
  2       0 0.04982206 0.72064057  0.2295374
```

1.8. Technical Details for Penalized Multicategory Logistic Regression

One of our versions of penalized multicategory logistic regression is similar to the regularized multinomial regression introduced by Friedman et al. (2010), the other one is a variation of the procedure of Zhu and Hastie (2004). An important difference is that we use a smooth approximation to the absolute value or norm function so that usual Newton-Raphson procedures (with step size correction) are applicable.

1.8.1. The Log-Likelihood-Function

Let $(\boldsymbol{X}, Y)$ be a random variable with values in $\mathbb{R}^d \times \{1, \ldots, L\}$ for some integer $L \geq 2$. We assume that

$$P(Y = y \,|\, \boldsymbol{X} = \boldsymbol{x}) \;=\; \exp(a_y + \boldsymbol{b}_y^\top \boldsymbol{x}) \Big/ \sum_{z=1}^{L} \exp(a_z + \boldsymbol{b}_z^\top \boldsymbol{x})$$

for unknown parameters $a_y \in \mathbb{R}$ and $\boldsymbol{b}_y \in \mathbb{R}^d$. For notational convenience we introduce the vectors

$$\boldsymbol{V} \;=\; \begin{pmatrix} 1 \\ \boldsymbol{X} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\theta}_y \;=\; \begin{pmatrix} a_y \\ \boldsymbol{b}_y \end{pmatrix}$$

in $\mathbb{R}^{d'}$ with $d' := 1 + d$. Then

$$P(Y = y \,|\, \boldsymbol{V} = \boldsymbol{v}) \;=\; \exp(\boldsymbol{\theta}_y^\top \boldsymbol{v}) \Big/ \sum_{z=1}^{L} \exp(\boldsymbol{\theta}_z^\top \boldsymbol{v}).$$

This parametrization is not unique, because $P(Y = y \,|\, \boldsymbol{V} = \boldsymbol{v})$ remains unchanged if we add one and the same arbitrary vector to all parameters $\boldsymbol{\theta}_y$. We will deal with this non-uniqueness later in various ways. Our goal is estimation of

$$\boldsymbol{\theta} \;=\; \left(\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top, \ldots, \boldsymbol{\theta}_L^\top\right)^\top \in \mathbb{R}^{Ld'},$$

based on independent data pairs $(\boldsymbol{V}_i, Y_i)$, $1 \le i \le n$, such that $\mathcal{L}(Y_i \,|\, \boldsymbol{V}_i) = \mathcal{L}(Y \,|\, \boldsymbol{V})$. Thus we consider the negative log-likelihood function

$$\Lambda(\boldsymbol{\theta}) \;:=\; \sum_{i=1}^{n} \Big(-\boldsymbol{\theta}_{Y_i}^{\top} \boldsymbol{V}_i + \log\Big(\sum_{y=1}^{L} \exp(\boldsymbol{\theta}_y^{\top} \boldsymbol{V}_i) \Big) \Big).$$

For the computation of the first and second derivatives of $\Lambda(\cdot)$ we shall use the following lemma:

Lemma 1.9. *Consider the functional*

$$\mathbb{R}^L \ni \boldsymbol{f} \;\mapsto\; \lambda(\boldsymbol{f}) := \log\Big(\sum_{y=1}^{L} \exp(f_y) \Big).$$

The gradient vector and Hessian matrix of this functional λ at $\boldsymbol{f}$ are given by

$$\boldsymbol{p}(\boldsymbol{f}) \;:=\; \Big(\exp(f_y) \Big/ \sum_{z=1}^{L} \exp(f_z) \Big)_{y=1}^{L}$$

and

$$\boldsymbol{h}(\boldsymbol{f}) \;:=\; \operatorname{diag}(\boldsymbol{p}(\boldsymbol{f})) - \boldsymbol{p}(\boldsymbol{f}) \boldsymbol{p}(\boldsymbol{f})^{\top},$$

respectively. Moreover, for any $\boldsymbol{v} \in \mathbb{R}^L$,

$$\boldsymbol{v}^{\top} \boldsymbol{h}(\boldsymbol{f}) \boldsymbol{v} \;\begin{cases} = \; 0 & \text{if } v_1 = v_2 = \cdots = v_L, \\ > \; 0 & \text{else.} \end{cases}$$

Proof. The formulae for gradient vector and Hessian matrix follow from elementary calculations. As to the sign of $\boldsymbol{v}^{\top} \boldsymbol{h}(\boldsymbol{f}) \boldsymbol{v}$, note that

$$\boldsymbol{v}^{\top} \boldsymbol{h}(\boldsymbol{f}) \boldsymbol{v} \;=\; \sum_{y=1}^{L} p_y(\boldsymbol{f}) v_y^2 - \Big(\sum_{z=1}^{L} p_z(\boldsymbol{f}) v_z \Big)^2 \;=\; \sum_{y=1}^{L} p_y(\boldsymbol{f}) (v_y - \bar{v}(\boldsymbol{f}))^2,$$

where $\bar{v}(\boldsymbol{f})$ stands for the weighted average $\sum_{y=1}^{L} p_y(\boldsymbol{f}) v_y$. Thus $\boldsymbol{v}^{\top} \boldsymbol{h}(\boldsymbol{f}) \boldsymbol{v}$ is non-negative and equals zero if, and only if, all components of $\boldsymbol{v}$ are identical. □

With Lemma 1.9 at hand one can easily determine the first and second derivatives of $\Lambda(\cdot)$. To formulate the results we use the Kronecker product

$\boldsymbol{B} \otimes \boldsymbol{C}$ of arbitrary matrices (or vectors) $\boldsymbol{B}$ and $\boldsymbol{C}$, namely

$$\boldsymbol{B} \otimes \boldsymbol{C} \;:=\; \begin{pmatrix} B_{11}\boldsymbol{C} & B_{12}\boldsymbol{C} & B_{13}\boldsymbol{C} & \cdots \\ B_{21}\boldsymbol{C} & B_{22}\boldsymbol{C} & B_{23}\boldsymbol{C} & \cdots \\ B_{31}\boldsymbol{C} & B_{32}\boldsymbol{C} & B_{33}\boldsymbol{C} & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix}.$$

For our purposes it is useful to know that

$$(\boldsymbol{B} \otimes \boldsymbol{C})^\top \;=\; \boldsymbol{B}^\top \otimes \boldsymbol{C}^\top \tag{1.22}$$

and

$$(\boldsymbol{B} \otimes \boldsymbol{C})(\boldsymbol{D} \otimes \boldsymbol{E}) \;=\; (\boldsymbol{BD}) \otimes (\boldsymbol{CE}) \tag{1.23}$$

for arbitrary matrices $\boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}, \boldsymbol{E}$ such that $\boldsymbol{BD}$ and $\boldsymbol{CE}$ are well-defined.

Moreover, for any dimension q, the standard basis of $\mathbb{R}^q$ is denoted by $\boldsymbol{e}_{q,1}, \boldsymbol{e}_{q,2}, \ldots, \boldsymbol{e}_{q,q}$.

Theorem 1.10. *Let $\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{v}) := (\boldsymbol{\theta}_y^\top \boldsymbol{v})_{y=1}^L$ for $\boldsymbol{v} \in \mathbb{R}^{d'}$. With $\boldsymbol{p}(\cdot)$ and $\boldsymbol{h}(\cdot)$ as in Lemma 1.9, the gradient vector and Hessian matrix of the negative log-likelihood $\Lambda(\cdot)$ at $\boldsymbol{\theta}$ are given by*

$$\boldsymbol{G}(\boldsymbol{\theta}) \;=\; \sum_{i=1}^n \big(\boldsymbol{p}(\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{V}_i)) - \boldsymbol{e}_{L,Y_i}\big) \otimes \boldsymbol{V}_i$$

and

$$\boldsymbol{H}(\boldsymbol{\theta}) \;=\; \sum_{i=1}^n \boldsymbol{h}(\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{V}_i)) \otimes (\boldsymbol{V}_i \boldsymbol{V}_i^\top),$$

respectively. The matrix $\boldsymbol{H}(\boldsymbol{\theta})$ is positive semidefinite. If the linear span of $\boldsymbol{V}_1, \boldsymbol{V}_2, \ldots, \boldsymbol{V}_n$ equals $\mathbb{R}^{d'}$, then for arbitrary $\boldsymbol{\delta} = (\boldsymbol{\delta}_1^\top, \boldsymbol{\delta}_2^\top, \ldots, \boldsymbol{\delta}_L^\top)^\top \in \mathbb{R}^{Ld'}$,

$$\boldsymbol{\delta}^\top \boldsymbol{H}(\boldsymbol{\theta}) \boldsymbol{\delta} \;=\; 0 \quad \textit{if, and only if,} \quad \boldsymbol{\delta}_1 = \boldsymbol{\delta}_2 = \cdots = \boldsymbol{\delta}_L.$$

Proof. Since $\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{v})$ is linear in $\boldsymbol{\theta}$, it follows from Lemma 1.9 that

$$\begin{aligned} &\lambda(\boldsymbol{f}(\boldsymbol{\theta} + \boldsymbol{\delta}, \boldsymbol{v})) - \lambda(\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{v})) \\ &\quad= \lambda\big(\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{v}) + \boldsymbol{f}(\boldsymbol{\delta}, \boldsymbol{v})\big) - \lambda(\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{v})) \\ &\quad= \boldsymbol{f}(\boldsymbol{\delta}, \boldsymbol{v})^\top \boldsymbol{p}(\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{v})) + 2^{-1} \boldsymbol{f}(\boldsymbol{\delta}, \boldsymbol{v})^\top \boldsymbol{h}(\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{v})) \boldsymbol{f}(\boldsymbol{\delta}, \boldsymbol{v}) + o\big(\|\boldsymbol{\delta}\|^2\big) \end{aligned}$$

as $\boldsymbol{\delta} \to \boldsymbol{0}$. Consequently,

$$\begin{aligned} \Lambda(\boldsymbol{\theta} + \boldsymbol{\delta}) - \Lambda(\boldsymbol{\theta}) = &\sum_{i=1}^n \big(-\boldsymbol{\delta}_{Y_i}^\top \boldsymbol{V}_i + \boldsymbol{f}(\boldsymbol{\delta}, \boldsymbol{V}_i)^\top \boldsymbol{p}(\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{V}_i))\big) \\ &+ 2^{-1} \sum_{i=1}^n \boldsymbol{f}(\boldsymbol{\delta}, \boldsymbol{V}_i)^\top \boldsymbol{h}(\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{V}_i)) \boldsymbol{f}(\boldsymbol{\delta}, \boldsymbol{V}_i) + o\big(\|\boldsymbol{\delta}\|^2\big). \end{aligned}$$

To obtain gradient and Hessian of $\Lambda(\cdot)$ explicitly, note that $\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{v}) = (\mathbb{I}_L \otimes \boldsymbol{v})^\top \boldsymbol{\theta} = (\mathbb{I}_L \otimes \boldsymbol{v}^\top)\boldsymbol{\theta}$. Thus the linear term (in $\boldsymbol{\delta}$) of the previous expansion of $\Lambda(\boldsymbol{\theta} + \boldsymbol{\delta})$ equals

$$\begin{aligned}
&\sum_{i=1}^{n} \big(\boldsymbol{f}(\boldsymbol{\delta}, \boldsymbol{V}_i)^\top \boldsymbol{p}(\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{V}_i)) - \boldsymbol{\delta}_{Y_i}^\top \boldsymbol{V}_i\big) \\
&\quad = \boldsymbol{\delta}^\top \sum_{i=1}^{n} \big((\mathbb{I}_L \otimes \boldsymbol{V}_i)\boldsymbol{p}(\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{V}_i)) - \boldsymbol{e}_{L,Y_i} \otimes \boldsymbol{V}_i\big) \\
&\quad = \boldsymbol{\delta}^\top \sum_{i=1}^{n} \big((\mathbb{I}_L \otimes \boldsymbol{V}_i)(\boldsymbol{p}(\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{V}_i)) \otimes 1) - \boldsymbol{e}_{L,Y_i} \otimes \boldsymbol{V}_i\big) \\
&\quad = \boldsymbol{\delta}^\top \sum_{i=1}^{n} \big(\boldsymbol{p}(\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{V}_i)) \otimes \boldsymbol{V}_i - \boldsymbol{e}_{L,Y_i} \otimes \boldsymbol{V}_i\big) \\
&\quad = \boldsymbol{\delta}^\top \sum_{i=1}^{n} \big(\boldsymbol{p}(\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{V}_i)) - \boldsymbol{e}_{L,Y_i}\big) \otimes \boldsymbol{V}_i,
\end{aligned}$$

while twice the quadratic term may be written as

$$\begin{aligned}
&\sum_{i=1}^{n} \boldsymbol{f}(\boldsymbol{\delta}, \boldsymbol{V}_i)^\top \boldsymbol{h}(\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{V}_i))\boldsymbol{f}(\boldsymbol{\delta}, \boldsymbol{V}_i) \\
&\quad = \boldsymbol{\delta}^\top \sum_{i=1}^{n} (\mathbb{I}_L \otimes \boldsymbol{V}_i)\boldsymbol{h}(\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{V}_i))(\mathbb{I}_L \otimes \boldsymbol{V}_i^\top)\, \boldsymbol{\delta} \\
&\quad = \boldsymbol{\delta}^\top \sum_{i=1}^{n} (\mathbb{I}_L \otimes \boldsymbol{V}_i)\big(\boldsymbol{h}(\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{V}_i)) \otimes 1\big)(\mathbb{I}_L \otimes \boldsymbol{V}_i^\top)\, \boldsymbol{\delta} \\
&\quad = \boldsymbol{\delta}^\top \Bigg(\sum_{i=1}^{n} \boldsymbol{h}(\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{V}_i)) \otimes (\boldsymbol{V}_i \boldsymbol{V}_i^\top)\Bigg)\boldsymbol{\delta}.
\end{aligned}$$

Finally, note that

$$\boldsymbol{\delta}^\top \boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{\delta} \;=\; \sum_{i=1}^{n} \boldsymbol{f}(\boldsymbol{\delta}, \boldsymbol{V}_i)^\top \boldsymbol{h}(\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{V}_i))\boldsymbol{f}(\boldsymbol{\delta}, \boldsymbol{V}_i) \;\geq\; 0$$

with equality if, and only if, $\boldsymbol{f}(\boldsymbol{\delta}, \boldsymbol{V}_i)^\top \boldsymbol{h}(\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{V}_i))\boldsymbol{f}(\boldsymbol{\delta}, \boldsymbol{V}_i) = 0$ for all indices i. According to Lemma 1.9, the latter condition is equivalent to

$$\boldsymbol{\delta}_1^\top \boldsymbol{V}_i = \boldsymbol{\delta}_2^\top \boldsymbol{V}_i = \cdots = \boldsymbol{\delta}_L^\top \boldsymbol{V}_i \quad \text{for } 1 \leq i \leq n.$$

But if the vectors $\boldsymbol{V}_i$ span the whole $\mathbb{R}^{d'}$, this is equivalent to all vectors $\boldsymbol{\delta}_y$ being identical. □

1.8.2. Regularizations

Regularization 0. One way to guarantee uniqueness of the parameter $\boldsymbol{\theta}$ is to require

$$\sum_{y=1}^{L} \boldsymbol{\theta}_y \;=\; \mathbf{0}.$$

More generally, let $\boldsymbol{\theta}_{[j]} := \big(\theta_{j,y}\big)_{y=1}^{L}$ with $\theta_{j,y}$ denoting the j-th component of θ_y. With $\mathbf{1}_L := (1,1,\ldots,1)^\top \in \mathbb{R}^L$, the previous condition means that

$$\mathbf{1}_L^\top \boldsymbol{\theta}_{[j]} \;=\; 0 \tag{1.24}$$

for all $j = 1, 2, \ldots, d'$. To enforce (1.24) at least for some j we can add

$$R_0(\boldsymbol{\theta}) \;:=\; 2^{-1} \sum_{j=1}^{d'} \sigma_j \big(\mathbf{1}_L^\top \boldsymbol{\theta}_{[j]}\big)^2$$

with a vector $\boldsymbol{\sigma} = (\sigma_j)_{j=1}^{d'} \in [0,\infty)^{d'}$ to $\Lambda(\theta)$. The choice of $\boldsymbol{\sigma}$ will depend on further regularization terms.

Theorem 1.11. *The gradient vector and Hessian matrix of R_0 at $\boldsymbol{\theta}$ are given by*

$$\boldsymbol{G}_{R,0}(\boldsymbol{\theta}) \;=\; \big((\mathbf{1}_L\mathbf{1}_L^\top) \otimes \operatorname{diag}(\boldsymbol{\sigma})\big)\boldsymbol{\theta} \quad \textit{and} \quad \boldsymbol{H}_{R,0}(\boldsymbol{\theta}) \;=\; (\mathbf{1}_L\mathbf{1}_L^\top) \otimes \operatorname{diag}(\boldsymbol{\sigma}),$$

respectively.

Proof. Expanding $R_0(\cdot)$ is rather simple, because it is a quadratic functional itself. Note first that

$$\begin{aligned} R_0(\boldsymbol{\theta}+\boldsymbol{\delta}) - R_0(\boldsymbol{\theta}) &= \sum_{j=1}^{d'} \sigma_j \big(\mathbf{1}_L^\top \boldsymbol{\delta}_{[j]}\big)\big(\mathbf{1}_L^\top \boldsymbol{\theta}_{[j]}\big) + 2^{-1} \sum_{j=1}^{d'} \sigma_j \big(\mathbf{1}_L^\top \boldsymbol{\delta}_{[j]}\big)^2 \\ &= \sum_{j=1}^{d'} \sigma_j \boldsymbol{\delta}_{[j]}^\top \mathbf{1}_L \mathbf{1}_L^\top \boldsymbol{\theta}_{[j]} + 2^{-1} \sum_{j=1}^{d'} \sigma_j \boldsymbol{\delta}_{[j]}^\top \mathbf{1}_L \mathbf{1}_L^\top \boldsymbol{\delta}_{[j]}. \end{aligned}$$

But the subvector $\boldsymbol{a}_{[j]}$ of $\boldsymbol{a} \in \{\boldsymbol{\theta}, \boldsymbol{\delta}\}$ may be written as $(\mathbb{I}_L \otimes \boldsymbol{e}_{d',j}^\top)\boldsymbol{a}$, so

$$\begin{aligned} \boldsymbol{\delta}_{[j]}^\top \mathbf{1}_L \mathbf{1}_L^\top \boldsymbol{a}_{[j]} &= \boldsymbol{\delta}^\top (\mathbb{I}_L \otimes \boldsymbol{e}_{d',j}) \mathbf{1}_L \mathbf{1}_L^\top (\mathbb{I}_L \otimes \boldsymbol{e}_{d',j}^\top) \boldsymbol{a} \\ &= \boldsymbol{\delta}^\top (\mathbb{I}_L \otimes \boldsymbol{e}_{d',j})((\mathbf{1}_L\mathbf{1}_L^\top) \otimes 1)(\mathbb{I}_L \otimes \boldsymbol{e}_{d',j}^\top)\boldsymbol{a} \\ &= \boldsymbol{\delta}^\top \big((\mathbf{1}_L\mathbf{1}_L^\top) \otimes (\boldsymbol{e}_{d',j}\boldsymbol{e}_{d',j}^\top)\big)\boldsymbol{a}, \end{aligned}$$

and $\sum_{j=1}^{d'} \sigma_j (\mathbf{1}_L\mathbf{1}_L^\top) \otimes (\boldsymbol{e}_{d',j}\boldsymbol{e}_{d',j}^\top) = (\mathbf{1}_L\mathbf{1}_L^\top) \otimes \operatorname{diag}(\boldsymbol{\sigma})$. Hence

$$R_0(\boldsymbol{\theta}+\boldsymbol{\delta}) - R_0(\boldsymbol{\theta}) = \boldsymbol{\delta}^\top \big((\mathbf{1}_L\mathbf{1}_L^\top) \otimes \operatorname{diag}(\boldsymbol{\sigma})\big)\,\boldsymbol{\theta} + 2^{-1}\boldsymbol{\delta}^\top \big((\mathbf{1}_L\mathbf{1}_L^\top) \otimes \operatorname{diag}(\boldsymbol{\sigma})\big)\,\boldsymbol{\delta}.$$

□

Regularization 1: penalizing subvectors. For logistic regression there are various good reasons to regularize the functional Λ or $\Lambda + R_0$. One is to avoid numerical problems. Another is to guarantee existence of a minimizer in cases where Λ alone has no minimizer. This happens if one subgroup $\{\boldsymbol{X}_i \colon Y_i = \theta_o\}$ is separated from $\{\boldsymbol{X}_i \colon Y_i \neq \theta_o\}$ by a hyperplane. Moreover, we want to favor parameter vectors with only few large components. A first way to do this would be to add the penalty

$$\sum_{j=1}^{d'} \tau_j \|\boldsymbol{\theta}_{[j]}\|$$

with $\boldsymbol{\tau} = (\tau_j)_{j=1}^{d'} \in [0,\infty)^{d'}$ to $\Lambda(\boldsymbol{\theta}) + R_0(\boldsymbol{\theta})$. Here and throughout, $\|\cdot\|$ denotes Euclidean norm. This regularization is motivated by Tibshirani's (1996) LASSO and similar in spirit to penalized logistic regression as proposed by Zhu and Hastie (2004). The latter authors used $\|\boldsymbol{\theta}_{[j]}\|^2$ instead of $\|\boldsymbol{\theta}_{[j]}\|$. To avoid problems with the non-smoothness of $\|\cdot\|$ at zero, we approximate it by a smooth function and consider

$$R_1(\boldsymbol{\theta}) \;:=\; \sum_{j=1}^{d'} \tau_j \bigl(\varepsilon^2 + \|\boldsymbol{\theta}_{[j]}\|^2\bigr)^{1/2}$$

for some small number $\varepsilon > 0$. Typically we consider $\tau_1 = 0$ and strictly positive parameters $\tau_2, \ldots, \tau_{d'}$. Note that $\|\boldsymbol{\theta}_{[j]} - c\,\mathbf{1}_L\|^2$ becomes minimal if c equals the mean $\mathbf{1}_L^\top \boldsymbol{\theta}_{[j]}/L$. Hence minimizing $\Lambda(\boldsymbol{\theta}) + R_0(\boldsymbol{\theta}) + R_1(\boldsymbol{\theta})$ enforces Condition (1.24) whenever $\sigma_j + \tau_j > 0$.

The following lemma is useful for the analysis of R_1:

Lemma 1.12. *Consider the functional*

$$\mathbb{R}^L \ni \boldsymbol{f} \;\mapsto\; \rho(\boldsymbol{f}) := \bigl(\varepsilon^2 + \|\boldsymbol{f}\|^2\bigr)^{1/2}.$$

The gradient vector and Hessian matrix of this functional ρ at $\boldsymbol{f}$ are given by

$$\boldsymbol{g}_\rho(\boldsymbol{f}) \;:=\; \rho(\boldsymbol{f})^{-1}\boldsymbol{f} \quad \textit{and} \quad \boldsymbol{h}_\rho(\boldsymbol{f}) \;:=\; \rho(\boldsymbol{f})^{-1}\mathbb{I}_L - \rho(\boldsymbol{f})^{-3}\boldsymbol{f}\boldsymbol{f}^\top,$$

respectively. Moreover, $\boldsymbol{h}_\rho(\boldsymbol{f})$ is positive definite for any $\boldsymbol{f} \in \mathbb{R}^L$.

PROOF. Since $(1+\delta)^{1/2} = 1 + \delta/2 - \delta^2/8 + O(\delta^3)$ as $\delta \to 0$,

$$\begin{aligned}\rho(\boldsymbol{f}+\boldsymbol{v}) &= \big(\varepsilon^2 + \|\boldsymbol{f}\|^2 + 2\boldsymbol{f}^\top\boldsymbol{v} + \|\boldsymbol{v}\|^2\big)^{1/2} \\ &= \rho(\boldsymbol{f})\Big(1 + \rho(\boldsymbol{f})^{-2}\big(2\boldsymbol{f}^\top\boldsymbol{v} + \|\boldsymbol{v}\|^2\big)\Big)^{1/2} \\ &= \rho(\boldsymbol{f}) + \rho(\boldsymbol{f})^{-1}\big(2\boldsymbol{f}^\top\boldsymbol{v} + \|\boldsymbol{v}\|^2\big)/2 - \rho(\boldsymbol{f})^{-3}\big(2\boldsymbol{f}^\top\boldsymbol{v} + \|\boldsymbol{v}\|^2\big)^2/8 \\ &\qquad + O(\|\boldsymbol{v}\|^3) \\ &= \rho(\boldsymbol{f}) + \rho(\boldsymbol{f})^{-1}\boldsymbol{f}^\top\boldsymbol{v} + \boldsymbol{v}^\top\big(\rho(\boldsymbol{f})^{-1}\mathbb{I} - \rho(\boldsymbol{f})^{-3}\boldsymbol{f}\boldsymbol{f}^\top\big)\boldsymbol{v}/2 + O(\|\boldsymbol{v}\|^3).\end{aligned}$$

This proves that gradient and Hessian of ρ at $\boldsymbol{f}$ are given by $\boldsymbol{g}_\rho(\boldsymbol{f}) := \rho(\boldsymbol{f})^{-1}\boldsymbol{f}$ and $\boldsymbol{h}_\rho(\boldsymbol{f}) := \rho(\boldsymbol{f})^{-1}\mathbb{I} - \rho(\boldsymbol{f})^{-3}\boldsymbol{f}\boldsymbol{f}^\top$, respectively. Moreover, since $\rho(\boldsymbol{f}) > \|\boldsymbol{f}\|$, it follows from the Cauchy-Schwarz inequality that

$$\boldsymbol{v}^\top\boldsymbol{h}_\rho(\boldsymbol{f})\boldsymbol{v} \;\geq\; \rho(\boldsymbol{f})^{-1}\big(1 - \rho(\boldsymbol{f})^{-2}\|\boldsymbol{f}\|^2\big)\|\boldsymbol{v}\|^2,$$

which is strictly positive for $\boldsymbol{v} \neq 0$. □

By means of Lemma 1.12 one can determine the first and second derivatives of the regularizing function $R_1(\cdot)$:

Theorem 1.13. *With $\boldsymbol{g}_\rho(\cdot)$ and $\boldsymbol{h}_\rho(\cdot)$ as in Lemma 1.12, the gradient vector and Hessian matrix of $R_1(\cdot)$ at $\boldsymbol{\theta}$ are given by*

$$\boldsymbol{G}_{R,1}(\boldsymbol{\theta}) = \sum_{j=1}^{d'} \tau_j\big(\boldsymbol{g}_\rho(\boldsymbol{\theta}_{[j]}) \otimes \boldsymbol{e}_{d',j}\big)$$

and

$$\boldsymbol{H}_{R,1}(\boldsymbol{\theta}) = \sum_{j=1}^{d'} \tau_j\big(\boldsymbol{h}_\rho(\boldsymbol{\theta}_{[j]}) \otimes (\boldsymbol{e}_{d',j}\boldsymbol{e}_{d',j}^\top)\big),$$

respectively. Moreover, for any $\boldsymbol{\delta} = (\boldsymbol{\delta}_1^\top, \ldots, \boldsymbol{\delta}_L^\top)^\top$ in $\mathbb{R}^{Ld'}$, $\boldsymbol{\delta}^\top\boldsymbol{H}_{R,1}(\boldsymbol{\theta})\boldsymbol{\delta} \geq 0$ with equality if, and only if,

$$\tau_1\boldsymbol{\delta}_{[1]} \;=\; \tau_2\boldsymbol{\delta}_{[2]} \;=\; \cdots \;=\; \tau_{d'}\boldsymbol{\delta}_{[d']} \;=\; \mathbf{0}.$$

PROOF. It follows from Lemma 1.12 that

$$R_1(\boldsymbol{\theta}+\boldsymbol{\delta}) - R_1(\boldsymbol{\theta}) \;=\; \sum_{j=1}^{d'} \tau_j\boldsymbol{\delta}_{[j]}^\top\boldsymbol{g}_\rho(\boldsymbol{\theta}_{[j]}) + 2^{-1}\sum_{j=1}^{d'} \tau_j\boldsymbol{\delta}_{[j]}^\top\boldsymbol{h}_\rho(\boldsymbol{\theta}_{[j]})\boldsymbol{\delta}_{[j]} + o\big(\|\boldsymbol{\delta}\|^2\big). \tag{1.25}$$

But $\boldsymbol{\delta}_{[j]} = (\mathbb{I}_L \otimes \boldsymbol{e}_{d',j})^\top \boldsymbol{\delta}$, whence

$$\begin{aligned}
\boldsymbol{\delta}_{[j]}^\top \boldsymbol{g}_\rho(\boldsymbol{\theta}_{[j]}) &= \boldsymbol{\delta}^\top (\mathbb{I}_L \otimes \boldsymbol{e}_{d',j})(\boldsymbol{g}_\rho(\boldsymbol{\theta}_{[j]}) \otimes 1) \\
&= \boldsymbol{\delta}^\top \big(\boldsymbol{g}_\rho(\boldsymbol{\theta}_{[j]}) \otimes \boldsymbol{e}_{d',j}\big), \\
\boldsymbol{\delta}_{[j]}^\top \boldsymbol{h}_\rho(\boldsymbol{\theta}_{[j]}) \boldsymbol{\delta}_{[j]} &= \boldsymbol{\delta}^\top (\mathbb{I}_L \otimes \boldsymbol{e}_{d',j})(\boldsymbol{h}_\rho(\boldsymbol{\theta}_{[j]}) \otimes 1)(\mathbb{I}_L \otimes \boldsymbol{e}_{d',j}^\top) \boldsymbol{\delta} \\
&= \boldsymbol{\delta}^\top \big(\boldsymbol{h}_\rho(\boldsymbol{\theta}_{[j]}) \otimes (\boldsymbol{e}_{d',j} \boldsymbol{e}_{d',j}^\top)\big) \boldsymbol{\delta}.
\end{aligned}$$

Plugging in the previous expressions within (1.25) yields the asserted expressions for gradient and Hessian.

The additional assertion about the Hessian matrix $\boldsymbol{H}_{R,1}(\boldsymbol{\theta})$ follows from (1.25) and the fact that all matrices $\boldsymbol{h}_\rho(\boldsymbol{\theta}_{[j]})$ are positive definite. □

Regularization 2: component-wise penalties. A simple form of regularization, analogous to Tibshirani's (1996) LASSO is to add the penalty

$$\sum_{j=1}^{d'} \tau_j \sum_{y=1}^{L} |\boldsymbol{\theta}_{j,y}| \;=\; \sum_{j=1}^{d'} \tau_j \|\theta_{[j]}\|_1$$

to $\Lambda(\boldsymbol{\theta}) + R_0(\boldsymbol{\theta})$. Again we use a smoothed version of this, namely

$$R_2(\boldsymbol{\theta}) \;:=\; \sum_{j=1}^{d'} \tau_j \sum_{y=1}^{L} (\varepsilon^2 + \theta_{j,y}^2)^{1/2}.$$

Application of Lemma 1.12 in the special case of $L = 1$ yields the derivatives

$$\rho'(f) \;=\; \frac{f}{(\varepsilon^2 + f^2)^{1/2}} \quad \text{and} \quad \rho''(f) \;=\; \frac{\varepsilon^2}{(\varepsilon^2 + f^2)^{3/2}}$$

of $\mathbb{R} \ni f \mapsto \rho(f) = (\varepsilon^2 + f^2)^{1/2}$. Hence the first two derivatives of R_2 have a rather simple form:

Theorem 1.14. *Let $\boldsymbol{v} = \boldsymbol{v}(\boldsymbol{\tau}, \boldsymbol{\theta}) \in \mathbb{R}^{Ld'}$ contain the vectors $\operatorname{diag}(\boldsymbol{\tau})\boldsymbol{\theta}_1$, $\operatorname{diag}(\boldsymbol{\tau})\boldsymbol{\theta}_2, \ldots, \operatorname{diag}(\boldsymbol{\tau})\boldsymbol{\theta}_L$ from top to bottom. The gradient vector and Hessian matrix of $R_2(\cdot)$ at $\boldsymbol{\theta}$ are given by*

$$\boldsymbol{G}_{R,2}(\boldsymbol{\theta}) = \operatorname{vec}\big((\tau_j \rho'(\theta_{j,y}))_{j \le d', y \le L}\big)$$

and

$$\boldsymbol{H}_{R,2}(\boldsymbol{\theta}) = \operatorname{diag}\Big(\operatorname{vec}\big((\tau_j \rho''(\theta_{j,y}))_{j \le d', y \le L}\big)\Big),$$

respectively.

In the previous theorem we use the notation $\operatorname{vec}(\boldsymbol{M})$ for a vector which is formed by stacking the columns of a matrix $\boldsymbol{M}$ (from left to right).

1.8.3. Strict Convexity and Coercivity

In this subsection we tackle the question when a unique minimizer of $\Lambda + R_0$ or of $\Lambda + R$ with $R = R_0 + R_1$ or $R_0 + R_2$ exists. Let us start with a general consideration: Suppose that $f : \mathbb{R}^q \to \mathbb{R}$ is continuously differentiable and convex. Then one can easily verify that the following three statements are equivalent:

$$\text{The set of minimizers of } f \text{ is nonvoid and compact;} \tag{1.26}$$

$$f \text{ is coercive, i.e. } \quad f(\boldsymbol{\theta}) \;\to\; \infty \quad \text{as } \|\boldsymbol{\theta}\| \to \infty; \tag{1.27}$$

$$\lim_{t\to\infty} \boldsymbol{\theta}^\top \nabla f(t\boldsymbol{\theta}) \;>\; 0 \quad \text{for any } \boldsymbol{\theta} \in \mathbb{R}^q \setminus \{\mathbf{0}\}. \tag{1.28}$$

The third statement (1.28) becomes more plausible when noting that $\mathbb{R} \ni t \mapsto \boldsymbol{\theta}^\top \nabla f(t\boldsymbol{\theta})$ is the derivative of the convex function $\mathbb{R} \ni t \mapsto f(t\boldsymbol{\theta})$.

Theorem 1.15. *Suppose that $\boldsymbol{V}_1, \boldsymbol{V}_2, \ldots, \boldsymbol{V}_n$ span $\mathbb{R}^{d'}$, and let $\boldsymbol{\sigma} \in (0,\infty)^{d'}$. Then the Hessian matrix of $\Lambda + R_0$ is positive definite everywhere. A (unique) minimizer of $\Lambda + R_0$ fails to exist if, and only if, there exist vectors $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_L \in \mathbb{R}^{d'}$ such that $\boldsymbol{\theta}_y \neq \boldsymbol{\theta}_z$ for some class labels $1 \le y < z \le L$ and*

$$\boldsymbol{\theta}_{Y_i}^\top \boldsymbol{V}_i \;=\; \max_{y=1,\ldots,L} \boldsymbol{\theta}_y^\top \boldsymbol{V}_i \quad \text{for } 1 \le i \le n. \tag{1.29}$$

Theorem 1.16. *Let $R = R_0 + R_1$ or $R = R_0 + R_2$. If $\sigma_1 > 0$ and $\tau_j > 0$ for $2 \le j \le d'$, then the Hessian matrix of $\Lambda + R$ is positive definite everywhere, and there exists a (unique) minimizer of $\Lambda + R$.*

Proof of Theorem 1.15. The Hessian of $\Lambda + R_0$ at $\boldsymbol{\theta}$ is the sum of the positive semidefinite matrices $\boldsymbol{H}(\boldsymbol{\theta})$ and $\boldsymbol{H}_{R,0}(\boldsymbol{\theta})$. According to Theorem 1.10, for any $\boldsymbol{\delta} \in \mathbb{R}^{Ld'}$, it follows from $\boldsymbol{\delta}^\top \boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{\delta} = 0$ that $\boldsymbol{\delta}_1 = \boldsymbol{\delta}_2 = \cdots = \boldsymbol{\delta}_L$. But $\boldsymbol{\delta}^\top \boldsymbol{H}_{R,0}(\boldsymbol{\theta})\boldsymbol{\delta} = 2R_0(\boldsymbol{\delta}) = \sum_{j=1}^{d'} \sigma_j (\mathbf{1}_L^\top \boldsymbol{\delta}_{[j]})^2 = 0$ implies that $\mathbf{1}_L^\top \boldsymbol{\delta}_{[j]} = 0$ for $1 \le j \le d'$, so $\sum_{y=1}^L \boldsymbol{\delta}_y = L\boldsymbol{\delta}_z = \mathbf{0}$ for $1 \le z \le L$.

As to the existence of a unique minimizer, note that strict convexity of $f = R + R_0$ implies that it has either a unique minimizer or no minimizer at all. Hence existence of a minimizer is equivalent to (1.28). Suppose that the latter condition is violated, i.e. $\lim_{t\to\infty} \boldsymbol{\theta}^\top \nabla f(t\boldsymbol{\theta}) \le 0$ for a fixed nonzero $\boldsymbol{\theta}$. Note that $\boldsymbol{\theta}^\top \nabla f(t\boldsymbol{\theta})$ is the sum of $\boldsymbol{\theta}^\top \boldsymbol{G}(t\boldsymbol{\theta})$ and $\boldsymbol{\theta}^\top \boldsymbol{G}_{R,0}(t\boldsymbol{\theta}) = 2tR_0(\boldsymbol{\theta})$. Moreover,

$$\begin{aligned} \boldsymbol{\theta}^\top \boldsymbol{G}(t\boldsymbol{\theta}) \;&=\; \sum_{i=1}^n \Big(\sum_{y=1}^L \frac{\exp(t\boldsymbol{\theta}_y^\top \boldsymbol{V}_i)}{\sum_{z=1}^L \exp(t\boldsymbol{\theta}_z^\top \boldsymbol{V}_i)} \boldsymbol{\theta}_y^\top \boldsymbol{V}_i - \boldsymbol{\theta}_{Y_i}^\top \boldsymbol{V}_i \Big) \\ &\to\; \sum_{i=1}^n \Big(\max_{y=1,\ldots,L} \boldsymbol{\theta}_y^\top \boldsymbol{V}_i - \boldsymbol{\theta}_{Y_i}^\top \boldsymbol{V}_i \Big), \end{aligned}$$

which is certainly nonnegative. Hence, our assumption entails that $R_0(\boldsymbol{\theta}) = 0$, i.e. $\sum_{y=1}^{L} \boldsymbol{\theta}_y = \mathbf{0}$ and (1.29). Since $\boldsymbol{\theta} \neq \mathbf{0}$ and $\sum_{y=1}^{L} \boldsymbol{\theta}_y = \mathbf{0}$, the subvectors $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_L$ cannot be all identical.

On the other hand, suppose that (1.29) holds for some $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_L^\top)^\top$ such that $\boldsymbol{\theta}_y \neq \boldsymbol{\theta}_z$ for some $1 \leq y < z \leq L$. These properties remain unchanged if we subtract $L^{-1} \sum_{y=1}^{L} \boldsymbol{\theta}_y$ from all subvectors $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_L$. But then we have a nonzero vector $\boldsymbol{\theta} \in \mathbb{R}^{Ld'}$ such that $\boldsymbol{\theta}^\top \boldsymbol{G}_{R,0}(t\boldsymbol{\theta}) = 0$ and $\boldsymbol{\theta}^\top \boldsymbol{G}(t\boldsymbol{\theta}) \to 0$ as $t \to \infty$. □

Proof of Theorem 1.16. Suppose that for some $\boldsymbol{\theta}, \boldsymbol{\delta} \in \mathbb{R}^{Ld'}$,

$$\boldsymbol{\delta}^\top \boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{\delta} \;=\; \boldsymbol{\delta}^\top \boldsymbol{H}_{R,0}(\boldsymbol{\theta})\boldsymbol{\delta} \;=\; \boldsymbol{\delta}^\top \boldsymbol{H}_{R,k}(\boldsymbol{\theta})\boldsymbol{\delta} \;=\; 0,$$

where $k = 1$ or $k = 2$. It follows from $\boldsymbol{\delta}^\top \boldsymbol{H}_{R,k}(\boldsymbol{\theta})\boldsymbol{\delta} = 0$ that $\delta_{j,y} = 0$ for $2 \leq j \leq d'$ and $1 \leq y \leq L$. But then $\boldsymbol{\delta}^\top \boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{\delta} = 0$ is equivalent to $\delta_{1,1} = \delta_{1,2} = \cdots = \delta_{1,L}$, because all vectors $\boldsymbol{V}_i$ have first component one. Hence it follows from $\boldsymbol{\delta}^\top \boldsymbol{H}_{R,0}(\boldsymbol{\theta})\boldsymbol{\delta} = \sigma_1(\mathbf{1}_L^\top \boldsymbol{\delta}_{[1]})^2 = 0$ that $\mathbf{1}_L^\top \boldsymbol{\delta}_{[1]} = L\delta_{1,y} = 0$ for $1 \leq y \leq L$.

For the existence of a unique minimizer we employ (1.28) again.

□

1.8.4. Some Comments on the Implementation in `pvclass`

Representations with matrices. For various reasons it is better to work with the parameter matrix $\underline{\boldsymbol{\theta}} = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_L] \in \mathbb{R}^{d' \times L}$. Then $\boldsymbol{\theta} = \mathrm{vec}(\underline{\boldsymbol{\theta}})$. In `R` the operator $\mathrm{vec}(\cdot)$ is implemented as `as.vector(·)`. Together with the augmented data matrix $\underline{\boldsymbol{V}} = [\boldsymbol{V}_1, \boldsymbol{V}_2, \ldots, \boldsymbol{V}_n]^\top$, one may write

$$\underline{\boldsymbol{F}} := \big[\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{V}_1), \boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{V}_2), \ldots, \boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{V}_n)\big]^\top \;=\; \underline{\boldsymbol{V}}\,\underline{\boldsymbol{\theta}},$$

and this simplifies various computations, e.g. the determination of

$$\underline{\boldsymbol{P}} := \big[\boldsymbol{p}(\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{V}_1)), \boldsymbol{p}(\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{V}_2)), \ldots, \boldsymbol{p}(\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{V}_n))\big]^\top,$$

considerably. Another trick to speed up computations is the well-known relation

$$\boldsymbol{a} \otimes \boldsymbol{b} \;=\; \mathrm{vec}(\boldsymbol{b}\boldsymbol{a}^\top)$$

for arbitrary (column) vectors $\boldsymbol{a}$ and $\boldsymbol{b}$. In particular, the gradient $\boldsymbol{G}(\boldsymbol{\theta})$ of Λ at $\boldsymbol{\theta}$ may be represented as

$$\begin{aligned}\boldsymbol{G}(\boldsymbol{\theta}) &= \sum_{i=1}^{n}\big(\boldsymbol{p}(\boldsymbol{f}(\boldsymbol{\theta},\boldsymbol{V}_i)) - \boldsymbol{e}_{L,Y_i}\big)\otimes\boldsymbol{V}_i \\ &= \mathrm{vec}\Big(\sum_{i=1}^{n}\boldsymbol{V}_i\,\big(\boldsymbol{p}(\boldsymbol{f}(\boldsymbol{\theta},\boldsymbol{V}_i)) - \boldsymbol{e}_{L,Y_i}\big)^\top\Big) \\ &= \mathrm{vec}\big(\underline{\boldsymbol{V}}^\top(\underline{\boldsymbol{P}} - \underline{\boldsymbol{E}})\big),\end{aligned}$$

where $\underline{\boldsymbol{E}} := [\boldsymbol{e}_{L,Y_1}, \boldsymbol{e}_{L,Y_2}, \ldots, \boldsymbol{e}_{L,Y_n}]^\top$. For the Hessian $\boldsymbol{H}(\boldsymbol{\theta})$ we also avoid the summation of n Kronecker products as follows: For $y, z \in \{1, 2, \ldots, L\}$,

$$\big(\boldsymbol{H}(\boldsymbol{\theta})_{ij}\big)_{(y-1)d'<i\le yd',(z-1)d'<j\le zd'} = \underline{\boldsymbol{V}}^\top\Big(\big((\delta_{yz}\boldsymbol{P}_y - \boldsymbol{P}_y\odot\boldsymbol{P}_z)\mathbf{1}_{d'}^\top\big)\odot\underline{\boldsymbol{V}}\Big),$$

where $\odot$ stands for componentwise multiplication, and $\boldsymbol{P}_1, \boldsymbol{P}_2, \ldots, \boldsymbol{P}_L$ are the columns of $\underline{\boldsymbol{P}}$.

Normalization of $\underline{\boldsymbol{F}}$. Having computed $\underline{\boldsymbol{F}} = (F_{iy})_{i\le n, y\le L}$, one can write $\underline{\boldsymbol{P}} = (P_{iy})_{i\le n, y\le L}$ as

$$P_{iy} = \exp(F_{iy})\Big/\sum_{z=1}^{L}\exp(F_{iz}),$$

and

$$\Lambda(\boldsymbol{\theta}) = \sum_{i=1}^{n}\Big(\log\Big(\sum_{z=1}^{L}\exp(F_{iz})\Big) - F_{i,Y_i}\Big).$$

These representations become problematic numerically if some components of $\underline{\boldsymbol{F}}$ become very large or if a whole row of $\underline{\boldsymbol{F}}$ consists of very small (negative) values. To circumvent these problems, we subtract from each row of $\underline{\boldsymbol{F}}$ its maximum. This does not affect the previous expressions for P_{iy} or $\Lambda(\boldsymbol{\theta})$.

Choice of $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$. In our implementations in `pvclass`, we use three versions of penalized logistic regression, specified by the parameters `pen.method` and τ_o:

`pen.method`	R	$\boldsymbol{\sigma}$	$\boldsymbol{\tau}$
`vectors`	$R_0 + R_1$	$\mathbf{1}_{d'}$	$(\tau_o S_j)_{j=1}^{d'}$
`simple`	$R_0 + R_2$	$\boldsymbol{e}_{d',1}$	$(\tau_o S_j)_{j=1}^{d'}$
`none`	R_0	$\mathbf{1}_{d'}$	–

Here S_j is the sample standard deviation (within groups) of the j-th components of the vectors $\boldsymbol{V}_i$.

Starting values. In the standard model with $\mathbb{P}(Y = y) = w_y$ and $\mathcal{L}(\boldsymbol{X} \,|\, Y = y) = \mathcal{N}_d(\boldsymbol{\mu}_y, \boldsymbol{\Sigma})$,

$$\mathbb{P}(Y = y \,|\, \boldsymbol{X} = \boldsymbol{x}) \;=\; \frac{\exp(\log w_y - \boldsymbol{\mu}_y^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_y/2 + \boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_y)}{\sum_{z=1}^{L} \exp(\log w_z - \boldsymbol{\mu}_z^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_z/2 + \boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_z)}.$$

Hence a possible starting point for iterative minimization algorithms is given by

$$\boldsymbol{\theta}_y^{(0)} \;=\; \boldsymbol{\theta}_y^{(*)} - \boldsymbol{M}^{(*)} \quad \text{with} \quad \boldsymbol{\theta}_y^{(*)} \;:=\; \begin{pmatrix} \log \widehat{w}_y - \widehat{\boldsymbol{\mu}}_y^\top \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\boldsymbol{\mu}}_y/2 \\ \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\boldsymbol{\mu}}_y \end{pmatrix},$$

where $\widehat{w}_y$, $\widehat{\boldsymbol{\mu}}_y$ and $\widehat{\boldsymbol{\Sigma}}$ are the usual parameter estimators in linear discriminant analysis, while $\boldsymbol{M}^{(*)} \in \mathbb{R}^{d'}$ is a centering vector depending on the type of regularization. Its first component equals the average of the components of $\boldsymbol{\theta}_{[1]}^{(*)}$. For $2 \le j \le d'$, the j-th component of $\boldsymbol{M}^{(*)}$ is either the mean (`pen.method = vectors`) or the median (`pen.method = simple`) of the components of $\boldsymbol{\theta}_{[j]}^{(*)}$.

Solution paths. Suppose that f_1, f_2 are two convex and twice continuously differentiable functionals on $\mathbb{R}^q$. Suppose further that for any $t > 0$, the functional $f_1 + t f_2$ is coercive with positive definite Hessian matrix $D^2 f_1 + t D^2 f_2$ everywhere. This entails that for each $t > 0$ there exists a unique minimizer $\boldsymbol{\theta}(t) \in \mathbb{R}^q$ of $f_1 + t f_2$; indeed $\boldsymbol{\theta}(t)$ is the unique solution $\boldsymbol{\theta}$ of the equation

$$\nabla f_1(\boldsymbol{\theta}) + t \nabla f_2(\boldsymbol{\theta}) \;=\; 0.$$

It follows from the implicit mapping theorem, applied to the function $\mathbb{R}^d \times \mathbb{R} \ni (\boldsymbol{x}, t) \mapsto \nabla f_1(\boldsymbol{x}) + t \nabla f_2(\boldsymbol{x}) \in \mathbb{R}^d$, that $(0, \infty) \ni t \mapsto \boldsymbol{\theta}(t)$ is also continuously differentiable with derivative

$$\boldsymbol{\theta}'(t) \;=\; -\big(D^2 f_1(\boldsymbol{\theta}(t)) + t D^2 f_2(\boldsymbol{\theta}(t))\big)^{-1} \nabla f_2(\boldsymbol{\theta}(t)).$$

These considerations are useful when minimizing $\Lambda + R_0 + \tau_o R_k$ for different values of $\tau_o > 0$. Having determined the minimizer $\boldsymbol{\theta}(\tau_o)$ for some value of τ_o, a good starting point for the Newton procedure with τ_* close to τ_o is given by

$$\begin{aligned} \boldsymbol{\theta}^{(0)} &:= \boldsymbol{\theta}(\tau_o) - (\tau_* - \tau_o)\big(\boldsymbol{H}(\boldsymbol{\theta}(\tau_o)) + \boldsymbol{H}_{R,0}(\boldsymbol{\theta}(\tau_o)) \\ &\qquad + \tau_o \boldsymbol{H}_{R,k}(\boldsymbol{\theta}(\tau_o))\big)^{-1} \boldsymbol{G}_{R,k}(\boldsymbol{\theta}(\tau_o)) \\ &= \boldsymbol{\theta}(\tau_o) - (\tau_*/\tau_o - 1)\big(\boldsymbol{H}(\boldsymbol{\theta}(\tau_o)) + \boldsymbol{H}_{R,0}(\boldsymbol{\theta}(\tau_o)) \\ &\qquad + \tau_o \boldsymbol{H}_{R,k}(\boldsymbol{\theta}(\tau_o))\big)^{-1} \big(\tau_o \boldsymbol{G}_{R,k}(\boldsymbol{\theta}(\tau_o))\big). \end{aligned}$$

2. Choice of Tuning Parameters

Some of the test statistics we use depend on a tuning parameter such as the k in the nearest neighbor method or the penalty parameter τ in the logistic regression. We want to choose them in a data-driven way which preserves the symmetry in $(\boldsymbol{X}_{I(j)})_{j=1}^{N_\theta}$.

Our first approach was to optimize the estimated expectations of the p-values. To choose the tuning parameter for the p-value $\pi_\theta(\boldsymbol{X}, \mathcal{D})$, we add the new observation $\boldsymbol{X}$ to the training data with class label θ. Then we search for the parameter which minimizes the sum of the cross-validated p-values

$$\sum_{i=1}^{n} \pi_\theta(\boldsymbol{X}_i, \mathcal{D}).$$

Unfortunately, this method chooses mostly small values for k or the regularization parameter. The reason for this is overfitting. It selects the parameter for which the classes are separated best. However, it is not taken into account how the p-values change, if the training data vary slightly. For example if we add a penalty term in the logistic regression, the separation of the training data gets worse, but we gain stability.

2.1. Stability

In a second approach we want to maximize the stability, i.e. $T_\theta(\boldsymbol{X}, \mathcal{D})$ should take big values for observations not belonging to class θ. To find the tuning parameter which maximizes the stability, we add the new observation $\boldsymbol{X}$ to the training data with class label θ. Then we compute for all training observations with $Y_i \neq \theta$ the test statistic

$$T_\theta^{(\tau)}(\boldsymbol{X}_i, \mathcal{D}_i(\boldsymbol{X}_i, \boldsymbol{X}, \theta)),$$

where $\mathcal{D}_i(\boldsymbol{X}_i, \boldsymbol{X}, \theta)$ denotes the training data after adding the observation $(\boldsymbol{X}, \theta)$ and setting the class label of observation $\boldsymbol{X}_i$ to θ. Then we take the sum of these values

$$S(\tau, \boldsymbol{X}, \theta) := \sum_{i:\, Y_i \neq \theta} T_\theta^{(\tau)}(\boldsymbol{X}_i, \mathcal{D}_i(\boldsymbol{X}_i, \boldsymbol{X}, \theta))$$

and search for the parameter τ^* which maximizes $S(\tau, \boldsymbol{X}, \theta)$:

$$\tau^*(\boldsymbol{X}, \theta) := \arg\max_{\tau} S(\tau, \boldsymbol{X}, \theta).$$

2.1.1. Subsampling

To determine the optimal tuning parameters for a new observation $\boldsymbol{X}$ and all potential class memberships, the test statistic has to be computed $(L-1)\cdot n \cdot l$ times, where l is the number of tuning parameters from which we want to choose the optimal one. This can be very computer-intensive, especially for penalized logistic regression in high dimensions. One way to reduce computation time is to take a random subset of the training data containing m observations per class $b \neq \theta$ and compute $T_\theta^{(\tau)}(\boldsymbol{X}_i, \mathcal{D}_i(\boldsymbol{X}_i, \boldsymbol{X}, \theta))$ only for these observations. Subsampling is particularly useful for large training sets.

2.1.2. Extended Golden Section Search

Another way to reduce computation time is to consider only few values for τ instead of a whole grid. We observed in simulated and real data examples of penalized logistic regression that $S(\tau, \boldsymbol{X}, \theta)$ is a unimodal function of τ for fixed $\boldsymbol{X}$ and θ, at least for reasonable intervals $[a, b]$.

Now suppose that $\tau^* \in [a, b]$ and $S(\tau)\colon [a, b] \to \mathbb{R}$ is *unimodal*, i.e. S is strictly increasing on $[a, \tau^*]$ and strictly decreasing on $[\tau^*, b]$. In this case we can use the golden section search.

Let $a \leq q < r < s < t \leq b$ such that $\tau^* \in [q, t]$. At the beginning of the algorithm we set $q = a$ and $t = b$. Then $S(r) \leq S(s)$ implies $\tau^* \in [r, t]$ and we replace $(q < r < s < t)$ by $(r < s < s' < t)$. In case of $S(r) > S(s)$ we know that $\tau^* \in [q, s]$ and replace $(q < r < s < t)$ by $(q < r' < r < s)$.

The new point r' or s', respectively, can be chosen in different ways. For the golden section search (Kiefer, 1953) we choose $r' = Cq + Bs$ or $s' = Br + Ct$, respectively, with $C = (\sqrt{5} - 1)/2 \approx 0.618$ and $B = 1 - C \approx 0.382$, i.e. we divide the interval in the golden ratio. In this way we can guarantee that the quadruples $(q < r < s < t), (r < s < s' < t)$ and $(q < r' < r < s)$ differ only by an affine transformation. Moreover, we get in each step a reduction of the interval length by the factor C. We stop the algorithm when $t - q \leq \delta$ and end up with an interval containing τ^* with length smaller than $C\delta$. Now we set $\tau^* = \arg\max(S(q), S(r), S(s), S(t))$. With this algorithm we don't get the exact argument of the maximum of S. However, the resulting p-values do not depend too sensitively on the exact choice of τ and it suffices to choose a τ which is not too far from the arg max. For the same reason, we can choose a rather large value for δ.

Since we don't know an interval $[a, b]$ which contains τ^*, we extend the golden section search. First we search for such an interval and then we start the

golden section search. As lower endpoint a we could take 0. However, to avoid numerical problems we suggest to take a small value greater than 0, e.g. 1. But a should be small enough, such that we may assume that $\tau^* \geq a$. To find the upper endpoint, we start with some point b and divide the interval $[a, b]$ in the golden ratio, i.e. $s = Ba + Cb$. If $S(s) \geq S(b)$, $\tau^* \in [a, b]$ and we can start the golden section search. Otherwise, τ^* could be greater than b and we define a new upper endpoint $b' = b + C(b - a)$. Then we iterate these steps until $S(s) \geq S(b)$. Pseudocode for the resulting algorithm is given in Algorithm 1. In the last step we assign $\tau^* = \arg\max(S(q), S_r, S_s)$, because $S(t) \leq S_s$ by construction of the algorithm.

Algorithm 1 $\tau^* \leftarrow$ **extendedGoldenSection**(S, a, b, δ)

```
(B, C) ← (0.382, 0.618)
s ← Ba + Cb
(S_s, S_b) ← (S(s), S(b))
if S_s < S_b then
  τ* ← extendedGoldenSection(S, a, b + C(b − a), δ)
  return τ*
end if
(q, r, t) ← (a, Ca + Bb, b)
S_r ← S(r)
while t − q > δ do
  if S_r > S_s then
    (r, s, t) ← (Cq + Bs, r, s)
    (S_r, S_s) ← (S(r), S_r)
  else
    (q, r, s) ← (r, s, Br + Ct)
    (S_r, S_s) ← (S_s, S(s))
  end if
end while
τ* ← arg max(S(q), S_r, S_s)
return τ*
```

2.2. Dimension Reduction

In very high-dimensional settings the computation of the p-values and the choice of τ, even with subsampling, become too computer-intensive. Therefore we reduce the dimension of the data before applying the procedure. The aim of this reduction is not to eliminate all noise variables, but to find a small enough subset of predictors for which our procedure is computationally feasible. We use l_1-penalized multicategory logistic regression to determine the

subset.

To preserve the symmetry in $(\boldsymbol{X}_{I(j)})_{j=1}^{N_\theta}$ we have two options. We can split the training set and use one part for reduction and the other as training set. If the training set is not big enough, we add the new observation to the training set and use the whole set for the reduction. The drawback of the latter method is that we have to do the reduction separately for each new observation and potential class membership.

2.3. Numerical Examples

2.3.1. Simulated Data

Example 2.1. Consider $L = 2$ classes with $P_\theta = \mathcal{N}_{100}(\boldsymbol{\mu}_\theta, \mathbb{I}_{100})$, where $\boldsymbol{\mu}_1 = (1, 0.5, 0.25, 0.125, 0, \ldots, 0)^\top$ and $\boldsymbol{\mu}_2 = -\boldsymbol{\mu}_1$. We simulated $N_1 = N_2 = 100$ training observations per class and computed $\pi_\theta^{(\tau)}(\boldsymbol{X}, \mathcal{D}(\boldsymbol{X}, \theta))$ for 100 test observations per class and $\tau = 1, 2, \ldots, 100$. Here $\pi_\theta^{(\tau)}$ denotes the p-value based on penalized logistic regression as described in Section 1.4. Figure 2.1 shows the rate of uniquely correct classified test observations for five different training and test sets. The missclassification rates depend heavily on the training set. Therefore we simulated 100 training and test sets and averaged the missclassification rates. The result is also shown in Figure 2.1 (bottom right). The corresponding distributions of τ^* are drawn as bar plots. Note that they are scaled to 1 and do not correspond to the scale of the y-axis. To determine the distribution of τ^* we used 20 training sets and 100 test observations per class and training set.

Example 2.2. Next we consider an example with $L = 5$ classes, $P_\theta = \mathcal{N}_{100}(\boldsymbol{\mu}_\theta, \mathbb{I}_{100})$ and $\boldsymbol{\mu}_\theta = (\mathbf{0}_{4(\theta-1)}, 2, 1, 0.5, 0.25, 0, \ldots, 0)^\top$, where $\mathbf{0}_j$ denotes the row vector with j zeros. We simulated $N_\theta = 100$ training observations and 40 test observations per class. Figure 2.2 shows the rate of uniquely correct classified test observations for five different training and test sets and averaged over 100 training and test sets (bottom right). The corresponding distributions of τ^* are drawn as bar plots. To determine the distribution of τ^* we used 5 training sets and 40 test observations per class and training set.

The choice of τ is good in both examples, but the procedure tends to pick slightly too large values for τ, i.e. to regularize slightly too strong. This is better than too little regularization, since a strong regularization avoids overfitting, which is a big problem, especially in high-dimensional settings.

Subsampling increases the variability of τ^*. But the results remain quite good for reasonable choices of m, e.g. $m = 10$.

Note that we averaged the rates of uniquely correct classified test observations over all classes, since these examples are symmetric.

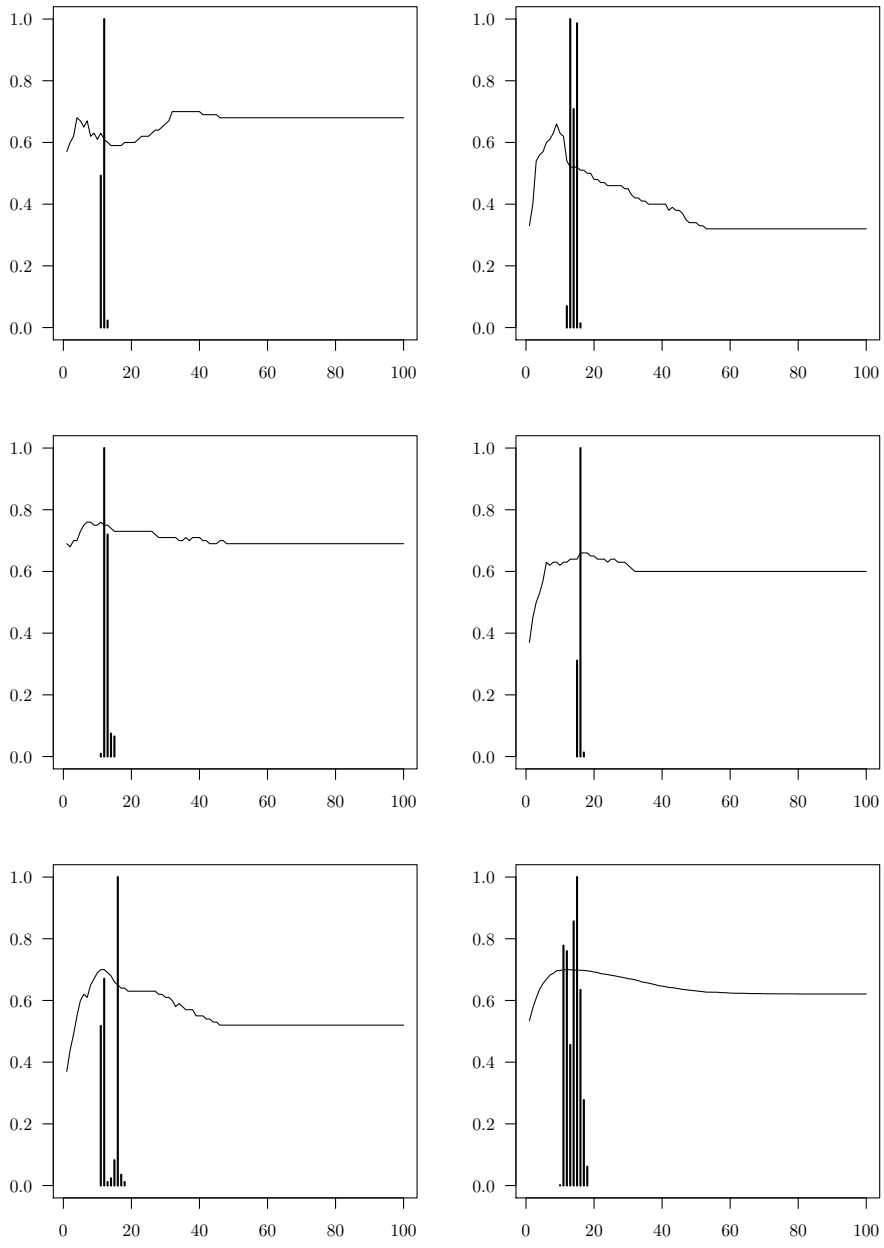

Figure 2.1.: Rates of uniquely correct classified test observations and distributions of τ^* for Example 2.1.

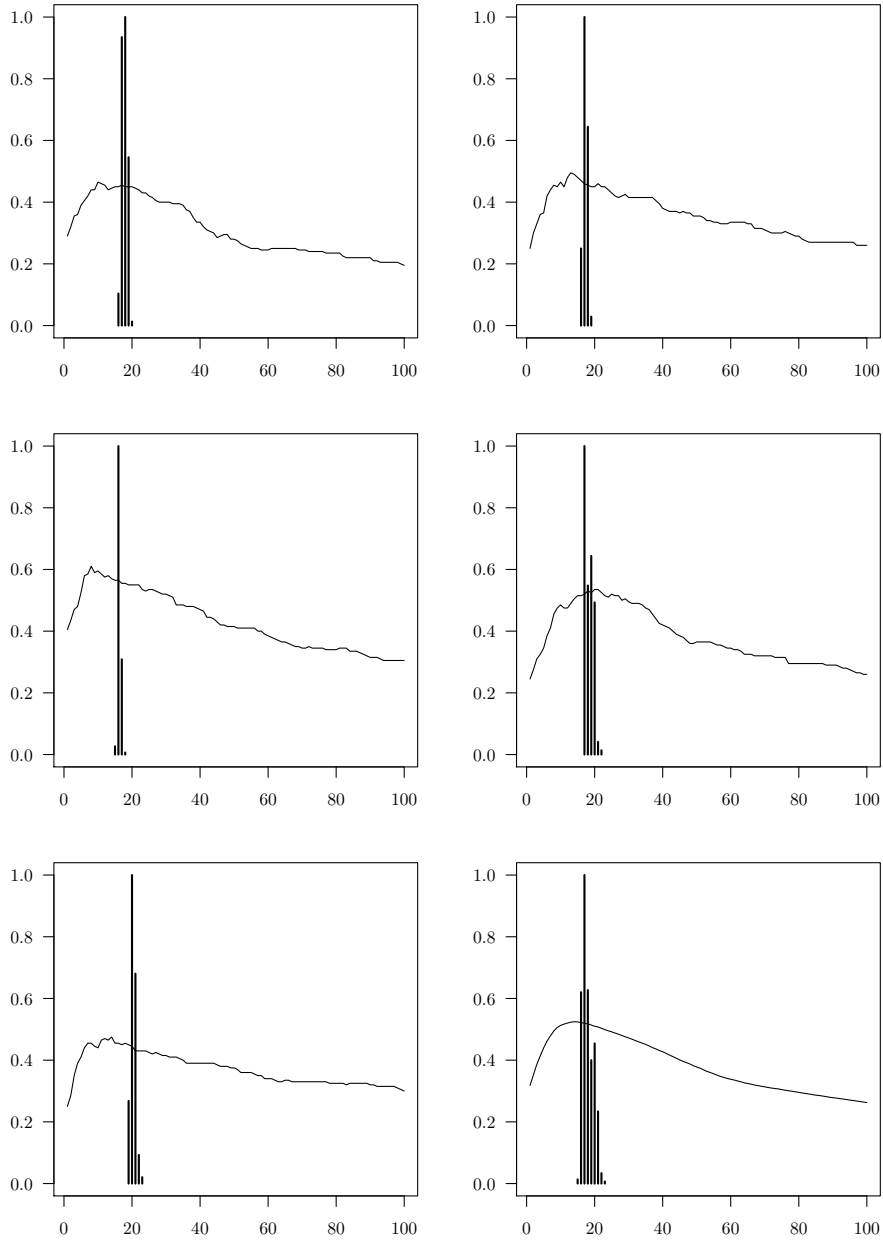

Figure 2.2.: Rates of uniquely correct classified test observations and distributions of τ^* for Example 2.2.

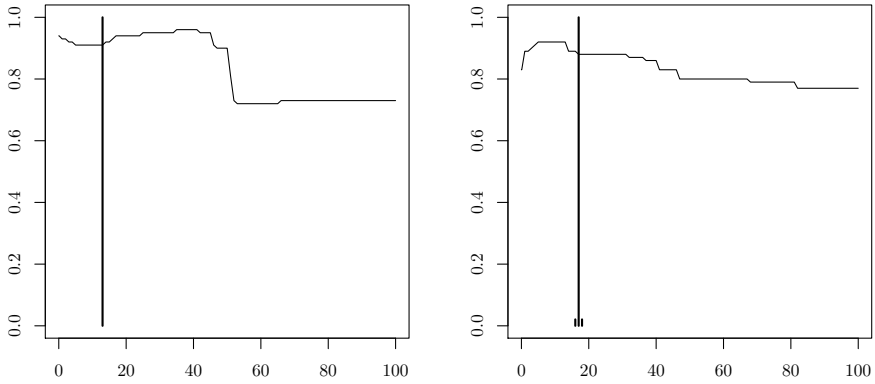

Figure 2.3.: Rates of uniquely correct classified test observations and distribution of τ^* for Example 2.3.

2.3.2. Real Data

Example 2.3 (Internet Ad)**.** We consider data of images on Internet pages provided by Kushmerick (1999). The binary response indicates whether the image is an advertisement. The preprocessed data set (Friedman et al., 2010) without missing values contains 1978 "nonads" and 381 ads. Three of the 1430 features are continuous, the others are binary. Computing p-values for such a high-dimensional problem would be too computer-intensive. Therefore we reduce the dimension as described in Section 2.2 using the R-package glmnet (Friedman et al., 2010). We split the data in a set for the reduction with 500 "nonads" and 100 ads, a test set with 100 observations per class and a training set with 1378 "nonads" and 181 ads.

For the reduction, we choose a penalty parameter for which we end up with a problem of dimension 120. Figure 2.3 shows the rate of uniquely correct classified test observations of class "nonads" (left) and ads (right). The distributions of τ^* are drawn as bar plots. They were determined using subsampling with $m = 100$.

Again, τ^* is not at the arg max of the rate of unique correct classification, but the rate for τ^* is near to the maximum. For class ads the procedure again regularizes slightly too strong.

Example 2.4 (Mushrooms)**.** The UCI Machine Learning Repository (Bache and Lichman, 2013) provides data of hypothetical samples corresponding to 23 species of gilled mushrooms. The binary response indicates whether the species is edible or poisonous. Most of the 22 categorical features describe the shape or the color. We removed 2 features. One has missing values and the other one takes only one value. After creating dummy-variables we end up

b	P(b,{})	P(b,{1})	P(b,{2})	P(b,{1,2})
1	0.07	0.93	0.00	0
2	0.08	0.00	0.92	0

Table 2.1.: Missclassification rates for $\tau = 40$ in Example 2.4.

with a problem of dimension 91. The data set contains observations of 4208 edible and 3916 poisonous mushrooms. We picked randomly 100 observations per class for the test set and used the rest as training set.

We computed the p-values for $\tau = 0.001, 0.01, 0.1, 1, 2, 3, \ldots, 100$. Table 2.1 shows the missclassification rates for $\tau = 40$. The choice of τ has no big influence on the result in this example. The rate of uniquely correct classified observations varies only between 0.9 and 0.94. All the p-values for the wrong classes are smaller than $\alpha = 0.05$. The two classes are perfectly separated, but for some of the observations both null hypotheses are rejected. They are located between the two class centers at a big distance from most of the training observations. Therefore both possible class memberships seem implausible for these observations.

Example 2.5 (Buerk)**.** For the hospital data described in Section 1.7 we computed p-values for $\tau = 1, 2, \ldots, 100$. The amount of regularization has no influence on the missclassification rates in this example.

3. Central Limit Theorems

In this chapter we derive two central limit theorems. First, we consider linear discriminant analysis and describe the asymptotic distribution of missclassification rates and cross-validated estimators thereof. Second, we consider p-values based on the plug-in statistic for the standard model and prove a central limit theorem for conditional inclusion probabilities and empirical conditional inclusion probabilities, which can be interpreted as estimators of the former.

We consider $L = 2$ classes with distributions P_1 and P_2 on $\mathcal{X} = \mathbb{R}^d$, which differ only by a shift. For notational convenience we assume without loss of generality that $Y_1 = 1$ and $Y_2 = 2$. Let $\mathbb{E}(\boldsymbol{X}_1) = \boldsymbol{\mu}_1$ and $\mathbb{E}(\boldsymbol{X}_2) = \boldsymbol{\mu}_2$ denote the mean vectors and suppose that

$$\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2.$$

Let $\tilde{\boldsymbol{X}}_i := \boldsymbol{X}_i - \boldsymbol{\mu}_{Y_i}$ denote the centered observation and $\boldsymbol{\Sigma} = \operatorname{Var}(\boldsymbol{X}_1) = \operatorname{Var}(\boldsymbol{X}_2)$ the common positive definite covariance matrix.

Moreover, let $P_0 := \mathcal{L}(\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_1 - \boldsymbol{\mu}_1)) = \mathcal{L}(\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_2 - \boldsymbol{\mu}_2)) = \mathcal{L}(\boldsymbol{Z})$ with $\boldsymbol{Z} := \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_1 - \boldsymbol{\mu}_1)$ and its first component Z_1.

We assume that $\boldsymbol{Z}$ has a differentiable Lebesgue density $f_{\boldsymbol{Z}} > 0$ and for all $\boldsymbol{U} \in \mathbb{R}^{d\times(d-1)}$ with $\boldsymbol{U}^\top \boldsymbol{U} = \mathbb{I}_{d-1}$

$$K(\boldsymbol{U}) := \int_{\mathbb{R}^{d-1}} \sup_{\boldsymbol{a}\in\boldsymbol{U}^\perp} \|\nabla f_{\boldsymbol{Z}}(\boldsymbol{a} + \boldsymbol{U}\boldsymbol{z})\|(1 + \|\boldsymbol{z}\|)^2 \,\mathrm{d}\boldsymbol{z} < \infty. \tag{3.1}$$

Here $\mathbb{I}_d$ denotes the d-dimensional identity matrix and $\boldsymbol{U}^\perp$ the orthogonal complement of the column space of $\boldsymbol{U}$. We define $\boldsymbol{\beta} := \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$, such that $\|\boldsymbol{\beta}\| = D_{\boldsymbol{\Sigma}}(\boldsymbol{\mu}_2, \boldsymbol{\mu}_1)$. Except for the proof of Lemma 3.13. There we define $\boldsymbol{\beta} := \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_\lambda - \boldsymbol{\mu}_\theta)$.

In this chapter we consider the class labels $Y_1, Y_2, \ldots, Y_n$ as fixed while $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ and $(\boldsymbol{X}, Y)$ are independent. We use $\widehat{w}_\theta := N_\theta/n$ as estimator for the prior weights w_θ. Note that the choice of $\widehat{w}_\theta$ is relevant only for the linear discriminant analysis. The p-values based on the plug-in statistic for the standard Gaussian model for two classes do not depend on the prior weights w_θ. Asymptotic statements are meant as

$$n \to \infty \quad \text{and} \quad \widehat{w}_\theta \to w_\theta \quad \text{for all } \theta \in \mathcal{Y}.$$

Convergence in probability is denoted by $\to_p$, convergence in law by $\to_{\mathcal{L}}$ and almost sure convergence by $\to_{\text{a.s.}}$. Generally we denote with $\widehat{\boldsymbol{v}}$ an estimator of the vector $\boldsymbol{v} = (v_1, v_2, \ldots, v_d)^\top$, by $\boldsymbol{\Delta}_{\boldsymbol{v}}$ the scaled difference $\sqrt{n}(\widehat{\boldsymbol{v}} - \boldsymbol{v}) = (\Delta_{\boldsymbol{v},1}, \Delta_{\boldsymbol{v},2}, \ldots, \Delta_{\boldsymbol{v},d})^\top$ and $\boldsymbol{Y}_{n,i}^{\boldsymbol{v}}$ is a summand depending on $\boldsymbol{X}_i$ such that $\boldsymbol{\Delta}_{\boldsymbol{v}} = \sum_{i=1}^n \boldsymbol{Y}_{n,i}^{\boldsymbol{v}} + \mathrm{o}_{\mathrm{p}}(1)$. Similarly, $\widehat{\boldsymbol{A}}$ is an estimator for the matrix

$$\boldsymbol{A} = \begin{pmatrix} A_{1,1} & \cdots & A_{1,d} \\ \vdots & \ddots & \vdots \\ A_{d,1} & \cdots & A_{d,d} \end{pmatrix},$$

$\boldsymbol{\Delta}_{\boldsymbol{A}}$ the scaled difference $\sqrt{n}(\widehat{\boldsymbol{A}} - \boldsymbol{A})$ and $\boldsymbol{Y}_{n,i}^{\boldsymbol{A}}$ is a summand depending on $\boldsymbol{X}_i$ such that $\boldsymbol{\Delta}_{\boldsymbol{A}} = \sum_{i=1}^n \boldsymbol{Y}_{n,i}^{\boldsymbol{A}} + \mathrm{o}_{\mathrm{p}}(1)$. Moreover, $\boldsymbol{v}_{i:j} = (v_i, v_{i+1}, \ldots, v_j)^\top$ is a vector consisting of the components i to j of $\boldsymbol{v}$.

The density of a random variable $\boldsymbol{\xi}$ is denoted by $f_{\boldsymbol{\xi}}$. If ξ is one-dimensional F_ξ denotes its distribution function.

For random matrices $\widehat{\boldsymbol{\Sigma}}$, $\operatorname{Var}(\widehat{\boldsymbol{\Sigma}}) := \operatorname{Var}(\operatorname{vec}(\widehat{\boldsymbol{\Sigma}}))$, where $\operatorname{vec}(\boldsymbol{M})$ denotes a vector which is formed by stacking the columns of a matrix $\boldsymbol{M}$ (from left to right). To formulate some of the results we use the Kronecker product $\otimes$ defined on page 29.

We denote the symmetric difference of two sets A and B with $A \triangle B := (A \setminus B) \cup (B \setminus A)$.

3.1. Half-Spaces

Let $\mathcal{H}$ denote the collection of all half-spaces in $\mathbb{R}^d$. A half-space is a set of the form

$$H(\boldsymbol{\beta}, \gamma) := \left\{\boldsymbol{z} \in \mathbb{R}^d \colon \boldsymbol{\beta}^\top \boldsymbol{z} + \gamma \leq 0\right\}$$

for $\boldsymbol{\beta} \in \mathbb{R}^d \setminus \{\boldsymbol{0}\}$ and $\gamma \in \mathbb{R}$.

The missclassification rates of the standard linear classifier can be written as a probability measure of a random subspace. Therefore we need some results about random half-spaces to describe asymptotic properties of missclassification rates. These results are also useful to describe the asymptotic behavior of inclusion probabilities for the p-values based on the plug-in statistic for the standard model.

3.1.1. Root-n-Consistency

The following lemma shows the root-n-consistency of the probability measure of random half-spaces under certain conditions.

Lemma 3.1. *Let $a \in \mathbb{R}$, $\boldsymbol{b} \in \mathbb{R}^d \setminus \{\boldsymbol{0}\}$ and $\boldsymbol{\psi} := (\boldsymbol{b}^\top, a)^\top$. Further let $\widehat{a}$ and $\widehat{\boldsymbol{b}}$ be random variables such that $\boldsymbol{\Delta}_{\boldsymbol{\psi}} := \sqrt{n}\begin{pmatrix}\widehat{\boldsymbol{b}} - \boldsymbol{b} \\ \widehat{a} - a\end{pmatrix} = \mathrm{O}_{\mathrm{p}}(1)$. Additionally let $\boldsymbol{X}$ be a random vector in $\mathbb{R}^d$ with measure P and differentiable density f which satisfies*

$$\|\mathbb{E}(\boldsymbol{X} \mid \boldsymbol{b}^\top \boldsymbol{X} = -a)\| < \infty \tag{3.2}$$

and for all $\boldsymbol{U} \in \mathbb{R}^{d\times(d-1)}$ with $\boldsymbol{U}^\top \boldsymbol{U} = \mathbb{I}_{d-1}$

$$K(\boldsymbol{U}) := \int\limits_{\mathbb{R}^{d-1}} \sup_{\boldsymbol{a}\in\boldsymbol{U}^\perp} \|\nabla f(\boldsymbol{a} + \boldsymbol{U}\boldsymbol{z})\|(1 + \|\boldsymbol{z}\|)^2 \,\mathrm{d}\boldsymbol{z} < \infty. \tag{3.3}$$

Then

$$\sqrt{n}P\big(\{\boldsymbol{x} \in \mathbb{R}^d \colon \widehat{\boldsymbol{b}}^\top \boldsymbol{x} + \widehat{a} \le 0\} \triangle \{\boldsymbol{x} \in \mathbb{R}^d \colon \boldsymbol{b}^\top \boldsymbol{x} + a \le 0\}\big) = \mathrm{O}_{\mathrm{p}}(1) \tag{3.4}$$

and

$$\begin{aligned}\sqrt{n}\big(&P\{\boldsymbol{x} \in \mathbb{R}^d \colon \widehat{\boldsymbol{b}}^\top \boldsymbol{x} + \widehat{a} \le 0\} - P\{\boldsymbol{x} \in \mathbb{R}^d \colon \boldsymbol{b}^\top \boldsymbol{x} + a \le 0\}\big) \\ &= \boldsymbol{c}^\top \boldsymbol{\Delta}_{\boldsymbol{\psi}} + \mathrm{O}_{\mathrm{p}}(n^{-1/2})\end{aligned} \tag{3.5}$$

with $\boldsymbol{c}^\top = -f_{\boldsymbol{b}^\top \boldsymbol{X}}(-a)\mathbb{E}\big((\boldsymbol{X}^\top, 1) \mid \boldsymbol{b}^\top \boldsymbol{X} = -a\big)$, where $f_{\boldsymbol{b}^\top \boldsymbol{X}}$ denotes the density of $\boldsymbol{b}^\top \boldsymbol{X}$.

Remark 3.2. The constant a in the previous lemma can be replaced by a deterministic convergent sequence $a_n \to a$. We will use the lemma with a sequence depending on $\widehat{w}_1/\widehat{w}_2$.

Remark 3.3. Condition (3.3) is satisfied for the multivariate t-Distribution with density

$$f(\boldsymbol{x}) = \det(\boldsymbol{\Sigma})^{-1/2} g_\nu\big((\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\big),$$

for a mean vector $\boldsymbol{\mu} \in \mathbb{R}^d$, a nonsingular covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$, $\nu > 1$ and $g_\nu(s) := C_\nu^{-1}(1 + s/\nu)^{-(\nu+d)/2}$ with some normalizing constant $C_\nu > 0$.

Proof of Remark 3.3. Without loss of generality let $\boldsymbol{\mu} = \boldsymbol{0}$ and $\boldsymbol{\Sigma} = \mathbb{I}_d$. Then

$$f(\boldsymbol{x}) = C_\nu^{-1}(1 + \nu^{-1}\|\boldsymbol{x}\|^2)^{-(\nu+d)/2}$$

and

$$\|\nabla f(\boldsymbol{x})\| = \frac{\nu + d}{C_\nu \nu}(1 + \nu^{-1}\|\boldsymbol{x}\|^2)^{-(\nu+d+2)/2}\|\boldsymbol{x}\|.$$

For $\boldsymbol{y} \in \arg\max_{\boldsymbol{x}\in\mathbb{R}^d} \|\nabla f(\boldsymbol{x})\|$ and $c := \|\boldsymbol{y}\|$,

$$\int_{\mathbb{R}^{d-1}} \sup_{\boldsymbol{a}\in\boldsymbol{U}^\perp} \|\nabla f(\boldsymbol{a}+\boldsymbol{U}\boldsymbol{z})\|(1+\|\boldsymbol{z}\|)^2 \mathbb{1}\{\|\boldsymbol{z}\| \le c\}\,\mathrm{d}\boldsymbol{z} < \infty.$$

If $\|\boldsymbol{z}\| > c$, $\|\nabla f(\boldsymbol{a}+\boldsymbol{U}\boldsymbol{z})\| \le \|\nabla f(\boldsymbol{U}\boldsymbol{z})\|$ for all $\boldsymbol{a} \in \boldsymbol{U}^\perp$. This implies for $c' > c$,

$$\begin{aligned}
&\int_{\mathbb{R}^{d-1}} \sup_{\boldsymbol{a}\in\boldsymbol{U}^\perp} \|\nabla f(\boldsymbol{a}+\boldsymbol{U}\boldsymbol{z})\|(1+\|\boldsymbol{z}\|)^2 \mathbb{1}\{c < \|\boldsymbol{z}\| \le c'\}\,\mathrm{d}\boldsymbol{z} \\
&\quad\le \int_{\mathbb{R}^{d-1}} \|\nabla f(\boldsymbol{U}\boldsymbol{z})\|(1+\|\boldsymbol{z}\|)^2 \mathbb{1}\{c < \|\boldsymbol{z}\| \le c'\}\,\mathrm{d}\boldsymbol{z} \\
&\quad= \frac{\nu+d}{C_\nu \nu} \int_{\mathbb{R}^{d-1}} (1+\nu^{-1}\|\boldsymbol{z}\|^2)^{-(\nu+d+2)/2} \|\boldsymbol{z}\|(1+\|\boldsymbol{z}\|)^2 \mathbb{1}\{c < \|\boldsymbol{z}\| \le c'\}\,\mathrm{d}\boldsymbol{z} \\
&\quad= \frac{(d-1)\tau_{d-1}(\nu+d)}{C_\nu \nu} \int_{\mathbb{R}} \frac{r(1+r)^2 r^{d-2}}{(1+\nu^{-1}r^2)^{(\nu+d+2)/2}} \mathbb{1}\{c < r \le c'\}\,\mathrm{d}\boldsymbol{z},
\end{aligned}$$

where $\tau_{d-1} := \pi^{(d-1)/2}\Gamma((d+1)/2)^{-1}$ is the volume of the $(d-1)$-dimensional unit sphere. By monotone convergence,

$$\begin{aligned}
&\int_{\mathbb{R}^{d-1}} \sup_{\boldsymbol{a}\in\boldsymbol{U}^\perp} \|\nabla f(\boldsymbol{a}+\boldsymbol{U}\boldsymbol{z})\|(1+\|\boldsymbol{z}\|)^2 \mathbb{1}\{c < \|\boldsymbol{z}\|\}\,\mathrm{d}\boldsymbol{z} \\
&\quad= \frac{(d-1)\tau_{d-1}(\nu+d)}{C_\nu \nu} \int_{\mathbb{R}} \frac{r(1+r)^2 r^{d-2}}{(1+\nu^{-1}r^2)^{(\nu+d+2)/2}} \mathbb{1}\{c < r\}\,\mathrm{d}r < \infty.
\end{aligned}$$

□

Proof of Lemma 3.1. Suppose that $\boldsymbol{b} = \boldsymbol{e}_1$, the first standard unit vector. Then

$$\begin{aligned}
&P\{\boldsymbol{x} \in \mathbb{R}^d \colon \widehat{\boldsymbol{b}}^\top \boldsymbol{x} + \widehat{a} \le 0\} - P\{\boldsymbol{x} \in \mathbb{R}^d \colon \boldsymbol{b}^\top \boldsymbol{x} + a \le 0\} \\
&\quad= P\Big\{ \Big(\boldsymbol{\psi} + n^{-\frac{1}{2}} \boldsymbol{\Delta}_{\boldsymbol{\psi}}\Big)^\top \begin{pmatrix} \boldsymbol{x} \\ 1 \end{pmatrix} \le 0 \Big\} - P\Big\{ \boldsymbol{\psi}^\top \begin{pmatrix} \boldsymbol{x} \\ 1 \end{pmatrix} \le 0 \Big\} \\
&\quad= P\Big\{ x_1 + a > 0 \ge x_1 + a + n^{-\frac{1}{2}} {\boldsymbol{\Delta}_{\boldsymbol{\psi}}}^\top \begin{pmatrix} \boldsymbol{x} \\ 1 \end{pmatrix} \Big\} \\
&\qquad - P\Big\{ x_1 + a \le 0 < x_1 + a + n^{-\frac{1}{2}} {\boldsymbol{\Delta}_{\boldsymbol{\psi}}}^\top \begin{pmatrix} \boldsymbol{x} \\ 1 \end{pmatrix} \Big\}
\end{aligned}$$

$$= P\Big\{x_1 + a > 0 \geq x_1(1 + n^{-\frac{1}{2}}\Delta_{\boldsymbol{\psi},1}) + a + n^{-\frac{1}{2}}\Big(\Delta_{\boldsymbol{\psi},d+1} + \sum_{i=2}^{d}\Delta_{\boldsymbol{\psi},i}x_i\Big)\Big\}$$
$$- P\Big\{x_1 + a \leq 0 < x_1(1 + n^{-\frac{1}{2}}\Delta_{\boldsymbol{\psi},1}) + a + n^{-\frac{1}{2}}\Big(\Delta_{\boldsymbol{\psi},d+1} + \sum_{i=2}^{d}\Delta_{\boldsymbol{\psi},i}x_i\Big)\Big\}.$$

With

$$\xi := \frac{-a - n^{-\frac{1}{2}}(\Delta_{\boldsymbol{\psi},d+1} + \sum_{i=2}^{d}\Delta_{\boldsymbol{\psi},i}x_i)}{1 + n^{-\frac{1}{2}}\Delta_{\boldsymbol{\psi},1}}$$

and the interval $I(\xi) := (\min(-a,\xi), \max(-a,\xi)]$ we may write

$$\begin{aligned}
&P\{\boldsymbol{x} \in \mathbb{R}^d \colon \widehat{\boldsymbol{b}}^\top \boldsymbol{x} + \widehat{a} \leq 0\} - P\{\boldsymbol{x} \in \mathbb{R}^d \colon \boldsymbol{b}^\top \boldsymbol{x} + a \leq 0\} \\
&\quad = P\{\xi \geq x_1 > -a\} - P\{\xi < x_1 \leq -a\} \\
&\quad = \int_{\mathbb{R}^{d-1}} \int_{I(\xi)} (-1)^{\mathbb{1}\{\xi < -a\}} f(\boldsymbol{x}) \,\mathrm{d}x_1 \,\mathrm{d}\boldsymbol{x}_{2:d} \\
&\quad = \int_{\mathbb{R}^{d-1}} \int_{I(\xi)} (-1)^{\mathbb{1}\{\xi < -a\}} \big(f(-a, \boldsymbol{x}_{2:d}) + (f(\boldsymbol{x}) - f(-a, \boldsymbol{x}_{2:d}))\big) \,\mathrm{d}x_1 \,\mathrm{d}\boldsymbol{x}_{2:d}.
\end{aligned}$$

Regarding the second summand,

$$\begin{aligned}
&\Big| \int_{\mathbb{R}^{d-1}} \int_{I(\xi)} (-1)^{\mathbb{1}\{\xi < -a\}} \big(f(\boldsymbol{x}) - f(-a, \boldsymbol{x}_{2:d})\big) \,\mathrm{d}x_1 \,\mathrm{d}\boldsymbol{x}_{2:d} \Big| \\
&\quad = \Big| \int_{\mathbb{R}^{d-1}} \int_{I(\xi)} (-1)^{\mathbb{1}\{\xi < -a\}} \int_{-a}^{x_1} \nabla f(\eta, \boldsymbol{x}_{2:d})^\top \boldsymbol{e}_1 \,\mathrm{d}\eta \,\mathrm{d}x_1 \,\mathrm{d}\boldsymbol{x}_{2:d} \Big| \\
&\quad \leq \int_{\mathbb{R}^{d-1}} \int_{I(\xi)} \int_{-a}^{x_1} \|\nabla f(\eta, \boldsymbol{x}_{2:d})\| \,\mathrm{d}\eta \,\mathrm{d}x_1 \,\mathrm{d}\boldsymbol{x}_{2:d} \\
&\quad \leq \int_{\mathbb{R}^{d-1}} \int_{I(\xi)} \int_{I(\xi)} \|\nabla f(\eta, \boldsymbol{x}_{2:d})\| \,\mathrm{d}\eta \,\mathrm{d}x_1 \,\mathrm{d}\boldsymbol{x}_{2:d} \\
&\quad = \int_{\mathbb{R}^{d-1}} \int_{I(\xi)} |\xi + a| \|\nabla f(\eta, \boldsymbol{x}_{2:d})\| \,\mathrm{d}\eta \,\mathrm{d}\boldsymbol{x}_{2:d}.
\end{aligned}$$

Cauchy-Schwarz inequality yields

$$\begin{aligned}|\xi + a| &= \frac{n^{-1/2}}{|1 + n^{-1/2}\Delta_{\boldsymbol{\psi},1}|} \left| \boldsymbol{\Delta}_{\boldsymbol{\psi}}^\top \begin{pmatrix} a \\ -\boldsymbol{x}_{2:d} \\ -1 \end{pmatrix} \right| \\ &\leq \frac{n^{-1/2}}{|1 + n^{-1/2}\Delta_{\boldsymbol{\psi},1}|} \|\boldsymbol{\Delta}_{\boldsymbol{\psi}}\| \sqrt{a^2 + 1} (1 + \|\boldsymbol{x}_{2:d}\|) \\ &= M_n (1 + \|\boldsymbol{x}_{2:d}\|)\end{aligned}$$

with

$$M_n := \frac{n^{-1/2} \|\boldsymbol{\Delta}_{\boldsymbol{\psi}}\| \sqrt{a^2 + 1}}{|1 + n^{-1/2}\Delta_{\boldsymbol{\psi},1}|} = \mathrm{O}_{\mathrm{p}}(n^{-1/2})$$

and thus

$$\begin{aligned}&\Big| \int_{\mathbb{R}^{d-1}} \int_{I(\xi)} (-1)^{\mathbb{1}\{\xi < -a\}} \big(f(\boldsymbol{x}) - f(-a, \boldsymbol{x}_{2:d})\big) \, \mathrm{d}x_1 \, \mathrm{d}\boldsymbol{x}_{2:d} \Big| \\ &\quad \leq M_n \int_{\mathbb{R}^{d-1}} \int_{I(\xi)} \|\nabla f(\eta, \boldsymbol{x}_{2:d})\| (1 + \|\boldsymbol{x}_{2:d}\|) \, \mathrm{d}\eta \, \mathrm{d}\boldsymbol{x}_{2:d} \\ &\quad \leq M_n \int_{\mathbb{R}^{d-1}} \int_{I(\xi)} \sup_{\tilde{\eta} \in I(\xi)} \|\nabla f(\tilde{\eta}, \boldsymbol{x}_{2:d})\| (1 + \|\boldsymbol{x}_{2:d}\|) \, \mathrm{d}\eta \, \mathrm{d}\boldsymbol{x}_{2:d} \\ &\quad \leq M_n^2 \int_{\mathbb{R}^{d-1}} \sup_{\tilde{\eta} \in I(\xi)} \|\nabla f(\tilde{\eta}, \boldsymbol{x}_{2:d})\| (1 + \|\boldsymbol{x}_{2:d}\|)^2 \, \mathrm{d}\boldsymbol{x}_{2:d} \\ &\quad \leq M_n^2 K = \mathrm{O}_{\mathrm{p}}(n^{-1})\end{aligned}$$

by Condition (3.3). Hence

$$\begin{aligned}&\sqrt{n}\big(P\{\boldsymbol{x} \in \mathbb{R}^d \colon \widehat{\boldsymbol{b}}^\top \boldsymbol{x} + \widehat{a} \leq 0\} - P\{\boldsymbol{x} \in \mathbb{R}^d \colon \boldsymbol{b}^\top \boldsymbol{x} + a \leq 0\}\big) \\ &\quad = \sqrt{n} \int_{\mathbb{R}^{d-1}} (a + \xi) f(-a, \boldsymbol{x}_{2:d}) \, \mathrm{d}\boldsymbol{x}_{2:d} + \mathrm{O}_{\mathrm{p}}(n^{-1/2}) \\ &\quad = \int_{\mathbb{R}^{d-1}} \frac{\Delta_{\boldsymbol{\psi},1} a - \Delta_{\boldsymbol{\psi},d+1} - \sum_{i=2}^d \Delta_{\boldsymbol{\psi},i} x_i}{1 + n^{-\frac{1}{2}} \Delta_{\boldsymbol{\psi},1}} f(-a, \boldsymbol{x}_{2:d}) \, \mathrm{d}\boldsymbol{x}_{2:d} + \mathrm{O}_{\mathrm{p}}(n^{-1/2}) \\ &\quad = \frac{\Delta_{\boldsymbol{\psi},1} a - \Delta_{\boldsymbol{\psi},d+1}}{1 + n^{-\frac{1}{2}} \Delta_{\boldsymbol{\psi},1}} \int_{\mathbb{R}^{d-1}} f(-a, \boldsymbol{x}_{2:d}) \, \mathrm{d}\boldsymbol{x}_{2:d} \\ &\qquad - \frac{1}{1 + n^{-\frac{1}{2}} \Delta_{\boldsymbol{\psi},1}} \sum_{i=2}^d \Delta_{\boldsymbol{\psi},i} \int_{\mathbb{R}^{d-1}} x_i f(-a, \boldsymbol{x}_{2:d}) \, \mathrm{d}\boldsymbol{x}_{2:d} + \mathrm{O}_{\mathrm{p}}(n^{-1/2})\end{aligned}$$

$$
\begin{aligned}
&= -\frac{1}{1 + n^{-\frac{1}{2}}\Delta_{\boldsymbol{\psi},1}} f_{X_1}(-a)\mathbb{E}\big((\boldsymbol{X}^\top, 1) \mid X_1 = -a\big)\boldsymbol{\Delta}_{\boldsymbol{\psi}} + \mathrm{O}_{\mathrm{p}}(n^{-1/2}) \\
&= -f_{X_1}(-a)\mathbb{E}\big((\boldsymbol{X}^\top, 1) \mid X_1 = -a\big)\boldsymbol{\Delta}_{\boldsymbol{\psi}} + \mathrm{O}_{\mathrm{p}}(n^{-1/2}).
\end{aligned}
$$

Now for an arbitrary vector $\boldsymbol{b} \in \mathbb{R}^d \setminus \{\boldsymbol{0}\}$, $\boldsymbol{b}^\top \boldsymbol{x} + a \leq 0$ if and only if $\boldsymbol{e}_1^\top \boldsymbol{y} + a/\|\boldsymbol{b}\| \leq 0$ with $\boldsymbol{y} = \boldsymbol{B}^\top \boldsymbol{x}$ and $\boldsymbol{B} = [\boldsymbol{b}/\|\boldsymbol{b}\|, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_d]$ such that $\boldsymbol{B}^\top \boldsymbol{B} = \mathbb{I}_d$. Further $\widehat{\boldsymbol{b}}^\top \boldsymbol{x} + \widehat{a} \leq 0$ if and only if $\widehat{\boldsymbol{b}'}^\top \boldsymbol{y} + \widehat{a}' \leq 0$ with $\widehat{a}' = \widehat{a}/\|\boldsymbol{b}\|$ and $\widehat{\boldsymbol{b}'} = \boldsymbol{B}^\top \widehat{\boldsymbol{b}}/\|\boldsymbol{b}\|$. Then with $\boldsymbol{V} := \boldsymbol{B}^\top \boldsymbol{X}$,

$$
\begin{aligned}
&P\{\boldsymbol{x} \in \mathbb{R}^d \colon \widehat{\boldsymbol{b}}^\top \boldsymbol{x} + \widehat{a} \leq 0\} - P\{\boldsymbol{x} \in \mathbb{R}^d \colon \boldsymbol{b}^\top \boldsymbol{x} + a \leq 0\} \\
&\quad = -f_{V_1}(-a/\|\boldsymbol{b}\|)\mathbb{E}\big((\boldsymbol{V}^\top, 1) \mid V_1 = -a/\|\boldsymbol{b}\|\big) \begin{pmatrix} \widehat{\boldsymbol{b}'} - \boldsymbol{e}_1 \\ \widehat{a}' - a/\|b\| \end{pmatrix} \\
&\quad = -f_{\boldsymbol{b}^\top \boldsymbol{X}/\|\boldsymbol{b}\|}(-a/\|\boldsymbol{b}\|)\mathbb{E}\big(((\boldsymbol{B}^\top \boldsymbol{X})^\top, 1) \mid \boldsymbol{b}^\top \boldsymbol{X} = -a\big) \\
&\qquad\qquad \cdot \frac{1}{\|b\|} \begin{pmatrix} \boldsymbol{B}^\top & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{b}} - \boldsymbol{b} \\ \widehat{a} - a \end{pmatrix} \\
&\quad = -f_{\boldsymbol{b}^\top \boldsymbol{X}}(-a)\mathbb{E}\big((\boldsymbol{X}^\top, 1) \mid \boldsymbol{b}^\top \boldsymbol{X} = -a\big) \begin{pmatrix} \widehat{\boldsymbol{b}} - \boldsymbol{b} \\ \widehat{a} - a \end{pmatrix}.
\end{aligned}
$$

Next we show assertion (3.4). Without loss of generality let $\boldsymbol{b} = \boldsymbol{e}_1$. Then

$$
\begin{aligned}
&P\big(\{\boldsymbol{x} \in \mathbb{R}^d \colon \widehat{\boldsymbol{b}}^\top \boldsymbol{x} + \widehat{a} \leq 0\} \triangle \{\boldsymbol{x} \in \mathbb{R}^d \colon \boldsymbol{b}^\top \boldsymbol{x} + a \leq 0\}\big) \\
&\quad = P\{\xi \geq x_1 > -a\} + P\{\xi < x_1 \leq -a\} \\
&\quad = \int\limits_{\mathbb{R}^{d-1}} \int\limits_{I(\xi)} f(\boldsymbol{x}) \, \mathrm{d}x_1 \, \mathrm{d}\boldsymbol{x}_{2:d} \\
&\quad = \int\limits_{\mathbb{R}^{d-1}} \int\limits_{I(\xi)} \big(f(-a, \boldsymbol{x}_{2:d}) + \big(f(\boldsymbol{x}) - f(-a, \boldsymbol{x}_{2:d})\big)\big) \, \mathrm{d}x_1 \, \mathrm{d}\boldsymbol{x}_{2:d} \\
&\quad = \int\limits_{\mathbb{R}^{d-1}} |a + \xi| f(-a, \boldsymbol{x}_{2:d}) \, \mathrm{d}\boldsymbol{x}_{2:d} + \mathrm{O}_{\mathrm{p}}(n^{-1}) \\
&\quad = \frac{1}{\sqrt{n}} \int\limits_{\mathbb{R}^{d-1}} \left| \frac{\Delta_{\boldsymbol{\psi},1} a - \Delta_{\boldsymbol{\psi},d+1} - \sum_{i=2}^d \Delta_{\boldsymbol{\psi},i} x_i}{1 + n^{-\frac{1}{2}}\Delta_{\boldsymbol{\psi},1}} \right| f(-a, \boldsymbol{x}_{2:d}) \, \mathrm{d}\boldsymbol{x}_{2:d} \\
&\qquad + \mathrm{O}_{\mathrm{p}}(n^{-1}).
\end{aligned}
$$

In the penultimate step we used that

$$\begin{aligned}
&\Big| \int\limits_{\mathbb{R}^{d-1}} \int\limits_{I(\xi)} \big(f(\boldsymbol{x}) - f(-a, \boldsymbol{x}_{2:d})\big)\, \mathrm{d}x_1\, \mathrm{d}\boldsymbol{x}_{2:d} \Big| \\
&\qquad \le \int\limits_{\mathbb{R}^{d-1}} \int\limits_{I(\xi)} \int\limits_{-a}^{x_1} \|\nabla f(\eta, \boldsymbol{x}_{2:d})\|\, \mathrm{d}\eta\, \mathrm{d}x_1\, \mathrm{d}\boldsymbol{x}_{2:d} \\
&\qquad \le M_n^2 K = \mathrm{O_p}(n^{-1}).
\end{aligned}$$

Thus

$$\begin{aligned}
&\sqrt{n} P\big(\{\boldsymbol{x} \in \mathbb{R}^d \colon \widehat{\boldsymbol{b}}^\top \boldsymbol{x} + \widehat{a} \le 0\} \triangle \{\boldsymbol{x} \in \mathbb{R}^d \colon \boldsymbol{b}^\top \boldsymbol{x} + a \le 0\}\big) \\
&\quad \le \left| \frac{\Delta_{\boldsymbol{\psi},1} a - \Delta_{\boldsymbol{\psi},d+1}}{1 + n^{-\frac{1}{2}} \Delta_{\boldsymbol{\psi},1}} \right| \int\limits_{\mathbb{R}^{d-1}} f(-a, \boldsymbol{x}_{2:d})\, \mathrm{d}\boldsymbol{x}_{2:d} \\
&\qquad + \sum_{i=2}^{d} \left(\left| \frac{\Delta_{\boldsymbol{\psi},i}}{1 + n^{-\frac{1}{2}} \Delta_{\boldsymbol{\psi},1}} \right| \Big| \int\limits_{\mathbb{R}^{d-1}} x_i f(-a, \boldsymbol{x}_{2:d})\, \mathrm{d}\boldsymbol{x}_{2:d} \Big| \right) + \mathrm{O_p}(n^{-1/2}) \\
&\quad = \mathrm{O_p}(1).
\end{aligned}$$

□

3.1.2. Empirical Processes

For the proofs of the central limit theorems we need some results about empirical processes. The *empirical measure* $\widehat{P}$ of a sample of independent random variables $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ in a measurable space $(\mathcal{X}, \mathcal{B})$ with distribution P is the discrete random measure given by

$$\widehat{P} := \frac{1}{n} \sum_{i=1}^{n} \delta_{\boldsymbol{X}_i},$$

with $\delta_{\boldsymbol{x}}$ denoting the Dirac measure at $\boldsymbol{x}$. For a measurable set $D \subset \mathcal{X}$,

$$\widehat{P}(D) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \in D\}.$$

For a probability measure P on $\mathbb{R}^d$ we consider the empirical process

$$\mathbb{B}_{P,n} = \mathbb{B}_{P,n}(H)_{H \in \mathcal{H}}$$

with

$$\mathbb{B}_{P,n}(H) := \sqrt{n}(\widehat{P} - P)(H),$$

where $\mathcal{H} := \{H(\boldsymbol{\beta},\gamma)\colon \boldsymbol{\beta} \in \mathbb{R}^d \setminus \{\mathbf{0}\}, \gamma \in \mathbb{R}\}$. Instead of $\mathbb{B}_{P,n}\big(H(\boldsymbol{\beta},\gamma)\big)$ we also write $\mathbb{B}_{P,n}(\boldsymbol{\beta},\gamma)$. For the proof of the following theorem we refer to van der Vaart and Wellner (1996). It is a consequence of the fact that $\mathcal{H}$ is a Vapnik–Červonenkis class.

Theorem 3.4. *The empirical process $\mathbb{B}_{P,n}$ converges in $l_\infty(\mathcal{H})$ weakly to a centered Gaussian process B_P with covariances*

$$\mathbb{E}\big(B_P(H)B_P(H')\big) = P(H \cap H') - P(H)P(H')$$

and uniformly continuous sample paths with respect to

$$\rho_P(H,H') := P(H \triangle H').$$

Moreover, $\mathcal{H}$ equipped with $\rho_P(H,H')$ is totally bounded and $\|\mathbb{B}_{P,n}\|_\infty = \mathrm{O}_\mathrm{p}(1)$, where $\|\cdot\|_\infty$ denotes the uniform norm.

Theorem 3.5. *Let $H(\boldsymbol{\beta},\gamma), H(\widehat{\boldsymbol{\beta}},\widehat{\gamma}) \in \mathcal{H}$ and P a probability measure on $\mathbb{R}^d$ such that $P\big(H(\boldsymbol{\beta},\gamma) \triangle H(\widehat{\boldsymbol{\beta}},\widehat{\gamma})\big) \to_p 0$. Then*

$$\mathbb{B}_{P,n}(\widehat{\boldsymbol{\beta}},\widehat{\gamma}) - \mathbb{B}_{P,n}(\boldsymbol{\beta},\gamma) \to_p 0.$$

Proof. For any fixed $\varepsilon > 0$, Theorem 3.4 implies that with asymptotic probability 1,

$$\begin{aligned}
|\mathbb{B}_n(\widehat{\boldsymbol{\beta}},\widehat{\gamma}) - \mathbb{B}_n(\boldsymbol{\beta},\gamma)| &\le \sup_{\substack{H(\tilde{\boldsymbol{\beta}},\tilde{\gamma})\in\mathcal{H}:\\ \rho_P(H(\boldsymbol{\beta},\gamma),H(\tilde{\boldsymbol{\beta}},\tilde{\gamma}))\le\varepsilon}} |\mathbb{B}_n(\tilde{\boldsymbol{\beta}},\tilde{\gamma}) - \mathbb{B}_n(\boldsymbol{\beta},\gamma)| \\
&\to_{\mathcal{L}} \sup_{\substack{H(\tilde{\boldsymbol{\beta}},\tilde{\gamma})\in\mathcal{H}:\\ \rho_P(H(\boldsymbol{\beta},\gamma),H(\tilde{\boldsymbol{\beta}},\tilde{\gamma}))\le\varepsilon}} |B_P(\tilde{\boldsymbol{\beta}},\tilde{\gamma}) - B_P(\boldsymbol{\beta},\gamma)| \\
&\to_p 0
\end{aligned}$$

as $\varepsilon \to 0$. □

3.2. Asymptotics of Estimators for Location and Scatter

The results of this section are not restricted to $L = 2$. For notational convenience we assume without loss of generality that $Y_1 = 1, Y_2 = 2, \ldots, Y_L = L$.

Lemma 3.6. *Let $\theta \in \mathcal{Y}$ and $\widehat{\boldsymbol{\mu}}_\theta = N_\theta^{-1} \sum_{i\in\mathcal{G}_\theta} \boldsymbol{X}_i$ be the standard estimator for $\boldsymbol{\mu}_\theta = \mathbb{E}(\boldsymbol{X}_i \mid Y_i = \theta)$. Suppose that $\mathbb{E}(\|\boldsymbol{X}\|^2) < \infty$. Then*

$$\boldsymbol{\Delta}_{\boldsymbol{\mu}_\theta} := \sqrt{n}(\widehat{\boldsymbol{\mu}}_\theta - \boldsymbol{\mu}_\theta) = \sum_{i=1}^n \boldsymbol{Y}_{n,i}^{\boldsymbol{\mu}_\theta} \to_{\mathcal{L}} \mathcal{N}_d(\mathbf{0}, w_\theta^{-1}\boldsymbol{\Sigma}),$$

where $\boldsymbol{Y}_{n,i}^{\boldsymbol{\mu}_\theta} := \sqrt{n}N_\theta^{-1}\mathbb{1}\{Y_i = \theta\}\tilde{\boldsymbol{X}}_i$.

Lemma 3.7. *Let* $\widehat{\boldsymbol{\Sigma}} := (n-L)^{-1}\sum_{\theta\in\mathcal{Y}}\sum_{i:\mathcal{G}_\theta}(\boldsymbol{X}_i-\widehat{\boldsymbol{\mu}}_\theta)(\boldsymbol{X}_i-\widehat{\boldsymbol{\mu}}_\theta)^\top$ *be the standard estimator for* $\boldsymbol{\Sigma}$ *and suppose that* $\mathbb{E}(\|\boldsymbol{X}\|^4)<\infty$. *Then*

$$\boldsymbol{\Delta}_{\boldsymbol{\Sigma}} := \sqrt{n}(\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}) = \sum_{i=1}^n \boldsymbol{Y}_{n,i}^{\boldsymbol{\Sigma}} + \mathrm{o_p}(1) \to_{\mathcal{L}} \mathcal{N}_{d\times d}(\mathbf{0}, \mathrm{Var}(\tilde{\boldsymbol{X}}_1\tilde{\boldsymbol{X}}_1^\top)) \tag{3.6}$$

and

$$\begin{aligned}\boldsymbol{\Delta}_{\boldsymbol{\Sigma}^{-1}} &:= \sqrt{n}(\widehat{\boldsymbol{\Sigma}}^{-1}-\boldsymbol{\Sigma}^{-1})) = \sum_{i=1}^n \boldsymbol{Y}_{n,i}^{\boldsymbol{\Sigma}^{-1}} + \mathrm{o_p}(1)\\ &\to_{\mathcal{L}} \mathcal{N}_{d\times d}(\mathbf{0}, (\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1})\,\mathrm{Var}(\tilde{\boldsymbol{X}}_1\tilde{\boldsymbol{X}}_1^\top)(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}))\end{aligned} \tag{3.7}$$

with $\boldsymbol{Y}_{n,i}^{\boldsymbol{\Sigma}} := n^{-1/2}(\tilde{\boldsymbol{X}}_i\tilde{\boldsymbol{X}}_i^\top-\boldsymbol{\Sigma})$ *and* $\boldsymbol{Y}_{n,i}^{\boldsymbol{\Sigma}^{-1}} := -\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}_{n,i}^{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}$.

Dümbgen et al. (2013) showed that the assumptions of the following lemma are satisfied for the M-estimators defined in Section 1.1.2 if $\mathcal{L}(\boldsymbol{X}\mid Y=\theta)$ is elliptically symmetric and $\mathbb{E}(\|\boldsymbol{X}\|^2)<\infty$.

Lemma 3.8. *Let*

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} &+ n^{-1}\sum_{\theta\in\mathcal{Y}}\sum_{i\in\mathcal{G}_\theta}\Big(g_\theta(\|\boldsymbol{\Sigma}^{-1/2}\tilde{\boldsymbol{X}}_i\|)\tilde{\boldsymbol{X}}_i\tilde{\boldsymbol{X}}_i^\top - h_\theta(\|\boldsymbol{\Sigma}^{-1/2}\tilde{\boldsymbol{X}}_i\|)\boldsymbol{\Sigma}\Big)\\ &+ \mathrm{o_p}(n^{-1/2})\end{aligned}$$

for continuous bounded functions g_θ, h_θ *such that* $g_\theta(r)r^2$ *is bounded for* $r\geq 0$. *Suppose that for* $Y_i=\theta$

$$\mathbb{E}\big(g_\theta(\|\boldsymbol{\Sigma}^{-1/2}\tilde{\boldsymbol{X}}_i\|)\tilde{\boldsymbol{X}}_i\tilde{\boldsymbol{X}}_i^\top - h_\theta(\|\boldsymbol{\Sigma}^{-1/2}\tilde{\boldsymbol{X}}_i\|)\boldsymbol{\Sigma}\big) = \mathbf{0}.$$

Then

$$\boldsymbol{\Delta}_{\boldsymbol{\Sigma}} = \sum_{i=1}^n \boldsymbol{Y}_{n,i}^{\boldsymbol{\Sigma}} + \mathrm{o_p}(1) \to_{\mathcal{L}} \mathcal{N}_{d\times d}(\mathbf{0}, \boldsymbol{V}) \tag{3.8}$$

and

$$\boldsymbol{\Delta}_{\boldsymbol{\Sigma}^{-1}} = \sum_{i=1}^n \boldsymbol{Y}_{n,i}^{\boldsymbol{\Sigma}^{-1}} + \mathrm{o_p}(1) \to_{\mathcal{L}} \mathcal{N}_{d\times d}(\mathbf{0}, (\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1})\boldsymbol{V}(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1})), \tag{3.9}$$

where

$$\begin{aligned}\boldsymbol{Y}_{n,i}^{\boldsymbol{\Sigma}} &:= n^{-1/2}\big(g_{Y_i}(\|\boldsymbol{\Sigma}^{-1/2}\tilde{\boldsymbol{X}}_i\|)\tilde{\boldsymbol{X}}_i\tilde{\boldsymbol{X}}_i^\top - h_{Y_i}(\|\boldsymbol{\Sigma}^{-1/2}\tilde{\boldsymbol{X}}_i\|)\boldsymbol{\Sigma}\big),\\ \boldsymbol{Y}_{n,i}^{\boldsymbol{\Sigma}^{-1}} &:= -\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}_{n,i}^{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}\end{aligned}$$

and

$$\boldsymbol{V} := \sum_{\theta\in\mathcal{Y}} w_\theta \operatorname{Var}\left(g_\theta(\|\boldsymbol{\Sigma}^{-1/2}\tilde{\boldsymbol{X}}_\theta\|)\tilde{\boldsymbol{X}}_\theta\tilde{\boldsymbol{X}}_\theta^\top - h_\theta(\|\boldsymbol{\Sigma}^{-1/2}\tilde{\boldsymbol{X}}_\theta\|)\boldsymbol{\Sigma}\right).$$

Proof of Lemma 3.6. We apply the central limit theorem (Theorem A.1) to the vectors $\boldsymbol{Y}_{n,i}^{\boldsymbol{\mu}_\theta}$. The assumptions are fulfilled, since $\mathbb{E}(\boldsymbol{Y}_{n,i}^{\boldsymbol{\mu}_\theta}) = \mathbf{0}$,

$$\sum_{i=1}^n \operatorname{Var}(\boldsymbol{Y}_{n,i}^{\boldsymbol{\mu}_\theta}) = N_\theta \operatorname{Var}(\boldsymbol{Y}_{n,\theta}^{\boldsymbol{\mu}_\theta}) = \frac{n}{N_\theta}\operatorname{Var}(\boldsymbol{X}_1) = \frac{1}{\widehat{w}_\theta}\boldsymbol{\Sigma} \to \frac{1}{w_\theta}\boldsymbol{\Sigma}$$

and by dominated convergence,

$$\begin{aligned}\sum_{i=1}^n \mathbb{E}\big(\|\boldsymbol{Y}_{n,i}^{\boldsymbol{\mu}_\theta}\|^2 \min(1, \|\boldsymbol{Y}_{n,i}^{\boldsymbol{\mu}_\theta}\|)\big) &= \frac{n}{N_\theta}\mathbb{E}\big(\|\tilde{\boldsymbol{X}}_\theta\|^2 \min(1, \sqrt{n}N_\theta^{-1}\|\tilde{\boldsymbol{X}}_\theta\|)\big)\\ &\to \mathbf{0}.\end{aligned}$$

□

Proof of Lemma 3.7. First note that

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} &= \frac{1}{n-L}\sum_{i=1}^n(\tilde{\boldsymbol{X}}_i\tilde{\boldsymbol{X}}_i^\top - \boldsymbol{\Sigma}) + \frac{L}{n-L}\boldsymbol{\Sigma}\\ &\quad - \sum_{\theta\in\mathcal{Y}}\frac{N_\theta}{n-L}\big(\widehat{\boldsymbol{\mu}}_\theta - \boldsymbol{\mu}_\theta\big)\big(\widehat{\boldsymbol{\mu}}_\theta - \boldsymbol{\mu}_\theta\big)^\top\\ &= n^{-1}\sum_{i=1}^n(\tilde{\boldsymbol{X}}_i\tilde{\boldsymbol{X}}_i^\top - \boldsymbol{\Sigma}) + \mathrm{O_p}(n^{-1}).\end{aligned}$$

We apply the central limit theorem to $\sum_{i=1}^n \boldsymbol{Y}_{n,i}^{\boldsymbol{\Sigma}} = \boldsymbol{\Delta}_{\boldsymbol{\Sigma}} + \mathrm{o_p}(1)$. As to the assumptions, note that $\mathbb{E}(\boldsymbol{Y}_{n,i}^{\boldsymbol{\Sigma}}) = \mathbf{0}$,

$$\sum_{i=1}^n \operatorname{Var}(\boldsymbol{Y}_{n,i}^{\boldsymbol{\Sigma}}) = n\operatorname{Var}(\boldsymbol{Y}_{n,1}^{\boldsymbol{\Sigma}}) = \operatorname{Var}\left(\tilde{\boldsymbol{X}}_1\tilde{\boldsymbol{X}}_1^\top\right)$$

and by dominated convergence,

$$\begin{aligned}&\sum_{i=1}^n \mathbb{E}\big(\|\boldsymbol{Y}_{n,i}^{\boldsymbol{\Sigma}}\|^2 \min(1, \|\boldsymbol{Y}_{n,i}^{\boldsymbol{\Sigma}}\|)\big)\\ &\quad= \mathbb{E}\big(\|\tilde{\boldsymbol{X}}_1\tilde{\boldsymbol{X}}_1^\top - \boldsymbol{\Sigma}\|^2 \min(1, n^{-1/2}\|\tilde{\boldsymbol{X}}_1\tilde{\boldsymbol{X}}_1^\top - \boldsymbol{\Sigma}\|)\big)\\ &\quad\to \mathbf{0}.\end{aligned}$$

For an invertible matrix $\boldsymbol{B} \in \mathbb{R}^{n\times n}$ and $\mathbb{R}^{n\times n} \ni \boldsymbol{\Delta} \to \mathbf{0}$, it is well-known (e.g. Taylor and Lay, 1980) that

$$(\boldsymbol{B}+\boldsymbol{\Delta})^{-1} = \boldsymbol{B}^{-1} - \boldsymbol{B}^{-1}\boldsymbol{\Delta}\boldsymbol{B}^{-1} + O(\|\boldsymbol{\Delta}\|^2).$$

Consequently,

$$\widehat{\boldsymbol{\Sigma}}^{-1} = \boldsymbol{\Sigma}^{-1} - n^{-1/2}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Delta}_{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1} + \mathrm{O}_{\mathrm{p}}(n^{-1})$$

and

$$\boldsymbol{\Delta}_{\boldsymbol{\Sigma}^{-1}} = -\boldsymbol{\Sigma}^{-1}\boldsymbol{\Delta}_{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1} + \mathrm{O}_{\mathrm{p}}(n^{-1/2}).$$

For arbitrary matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ such that $\boldsymbol{ABC}$ is well-defined, the relation

$$\operatorname{vec}(\boldsymbol{ABC}) = (\boldsymbol{C}^\top \otimes \boldsymbol{A}) \operatorname{vec}(\boldsymbol{B}) \tag{3.10}$$

is well-known and yields

$$\operatorname{vec}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Delta}_{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}\right) = (\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}) \operatorname{vec}\left(\boldsymbol{\Delta}_{\boldsymbol{\Sigma}}\right).$$

Now with $\boldsymbol{Y}_{n,i}^{\boldsymbol{\Sigma}^{-1}} := -\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}_{n,i}^{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}$, Claim (3.7) follows from (3.6). □

Proof of Lemma 3.8. By assumption, $\mathbb{E}(\boldsymbol{Y}_{n,i}^{\boldsymbol{\Sigma}}) = \mathbf{0}$. Moreover, since $\sqrt{n}\|\boldsymbol{Y}_{n,i}^{\boldsymbol{\Sigma}}\|$ is uniformly bounded,

$$\begin{aligned}
\sum_{i=1}^{n} \operatorname{Var}(\boldsymbol{Y}_{n,i}^{\boldsymbol{\Sigma}}) &= \sum_{\theta\in\mathcal{Y}} N_\theta \operatorname{Var}(\boldsymbol{Y}_{n,\theta}^{\boldsymbol{\Sigma}}) \\
&= \sum_{\theta\in\mathcal{Y}} \frac{N_\theta}{n} \operatorname{Var}\left(g_\theta(\|\boldsymbol{\Sigma}^{-1/2}\tilde{\boldsymbol{X}}_\theta\|)\tilde{\boldsymbol{X}}_\theta\tilde{\boldsymbol{X}}_\theta^\top - h_\theta(\|\boldsymbol{\Sigma}^{-1/2}\tilde{\boldsymbol{X}}_\theta\|)\boldsymbol{\Sigma}\right) \\
&\to \sum_{\theta\in\mathcal{Y}} w_\theta \operatorname{Var}\left(g_\theta(\|\boldsymbol{\Sigma}^{-1/2}\tilde{\boldsymbol{X}}_\theta\|)\tilde{\boldsymbol{X}}_\theta\tilde{\boldsymbol{X}}_\theta^\top - h_\theta(\|\boldsymbol{\Sigma}^{-1/2}\tilde{\boldsymbol{X}}_\theta\|)\boldsymbol{\Sigma}\right)
\end{aligned}$$

and for some constant $M < \infty$

$$\sum_{i=1}^{n} \mathbb{E}\big(\|\boldsymbol{Y}_{n,i}^{\boldsymbol{\Sigma}}\|^2 \min(1, \|\boldsymbol{Y}_{n,i}^{\boldsymbol{\Sigma}}\|)\big) \le n\mathbb{E}\big(n^{-1}M^2 \cdot \min(1, n^{-1/2}M)\big) \to 0.$$

Now the central limit theorem yields (3.8). Claim (3.9) can be proved the same way as Claim (3.7) in Lemma 3.7. □

3.3. A Central Limit Theorem for Missclassification Rates

In this section we examine the asymptotic properties of the standard linear classifier (1.3) for two classes, without assuming Gaussian distributions. More precisely, we describe the asymptotic distribution of the missclassification rates

$$R_\theta = \mathbb{P}\big(\widehat{Y}(\boldsymbol{X}, \mathcal{D}) \neq Y \mid Y = \theta, \mathcal{D}\big)$$

and the cross-validated estimators

$$\widehat{R}_\theta = N_\theta^{-1} \#\{i \in \mathcal{G}_\theta \colon \widehat{Y}(\boldsymbol{X}_i, \mathcal{D}_i) \neq Y_i\}$$

thereof.

Since we don't want to make assumptions on the convergence rate of $\widehat{w}_\theta$, we consider a reweighted version of the optimal classifier depending on $\widehat{w}_\theta$ instead of w_θ, namely

$$\widehat{Y}_n^*(\boldsymbol{x}) := \begin{cases} 1, & (\boldsymbol{x} - \boldsymbol{\mu}_{1,2})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) + \log(\widehat{w}_2/\widehat{w}_1) \leq 0 \\ 2, & (\boldsymbol{x} - \boldsymbol{\mu}_{1,2})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) + \log(\widehat{w}_2/\widehat{w}_1) > 0. \end{cases}$$

Note that we consider the class labels as fixed and therefore $\widehat{w}_\theta$ is deterministic and converges to w_θ by assumption.

We define $\gamma := \|\boldsymbol{\beta}\|/2 + \log(w_1/w_2)/\|\boldsymbol{\beta}\|$, $\boldsymbol{\nu} := \mathbb{E}(\boldsymbol{Z} \mid \widehat{Y}^*(\boldsymbol{X}_1) = 1)$, $\boldsymbol{u} := \|\boldsymbol{\beta}\|^{-1}\boldsymbol{\beta}$ and suppose that

$$\big\|\mathbb{E}\big(\boldsymbol{X}_1 \mid (\boldsymbol{X}_1 - \boldsymbol{\mu}_{1,2})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) + \log(\widehat{w}_2/\widehat{w}_1) = 0\big)\big\| < \infty, \tag{3.11}$$

where we condition on $\boldsymbol{X}_1$ being in the separating hyperplane of the optimal classifier.

The following two central limit theorems imply that, under certain conditions, the standard linear classifier for two classes is asymptotically optimal. Moreover, $\widehat{R}_1$ is a root-n-consistent estimator of R_1.

First, we consider the standard estimator of $\boldsymbol{\Sigma}$. In this case, the only assumptions we need to make about the distributions are a finite fourth moment and that (3.1) and (3.11) hold.

Theorem 3.9. *Let $L = 2$ and $\widehat{\boldsymbol{\Sigma}}$ be the standard estimator. Suppose that $\mathbb{E}(\|\boldsymbol{X}\|^4) < \infty$, (3.1) and (3.11) hold. Then for the standard linear classifier,*

$$\sqrt{n}\begin{pmatrix} R_1 - \mathbb{P}\big(\widehat{Y}_n^*(\boldsymbol{X}) \neq Y \;\big|\; Y = 1\big) \\ \widehat{R}_1 - \mathbb{P}\big(\widehat{Y}_n^*(\boldsymbol{X}) \neq Y \;\big|\; Y = 1\big) \end{pmatrix} \to_{\mathcal{L}} \mathcal{N}_2\big(\boldsymbol{0}, \boldsymbol{\Lambda}\big)$$

and the components of the covariance matrix $\boldsymbol{\Lambda}$ *are given by*

$$\begin{aligned}\Lambda_{1,1} &= \|\boldsymbol{\beta}\|^{-2} f_{\boldsymbol{u}^\top \boldsymbol{Z}}(\gamma)^2 \Big[\operatorname{Var}\big(\boldsymbol{\beta}^\top \boldsymbol{Z}(\boldsymbol{\nu} - 2^{-1}\boldsymbol{\beta})^\top \boldsymbol{Z}\big) + w_1^{-1}\|\boldsymbol{\nu}\|^2 \\ &\qquad + w_2^{-1}\|\boldsymbol{\nu} - \boldsymbol{\beta}\|^2 + 2\mathbb{E}\big((\boldsymbol{\beta}^\top \boldsymbol{Z})^2(\boldsymbol{\nu} - 2^{-1}\boldsymbol{\beta})^\top \boldsymbol{Z}\big)\Big], \\ \Lambda_{2,2} &= \Lambda_{1,1} + w_1^{-1}\big(\mathbb{P}(\widehat{Y}^*(\boldsymbol{X}_1) = 1) - \mathbb{P}(\widehat{Y}^*(\boldsymbol{X}_1) = 1)^2\big) \\ &\quad + 2\|\boldsymbol{\beta}\|^{-1} f_{\boldsymbol{u}^\top \boldsymbol{Z}}(\gamma) \operatorname{Cov}\big(\boldsymbol{\beta}^\top \boldsymbol{Z}(\boldsymbol{\nu} - 2^{-1}\boldsymbol{\beta})^\top \boldsymbol{Z}, \mathbb{1}\{\widehat{Y}^*(\boldsymbol{X}_1) = 1\}\big) \\ &\quad + 2w_1^{-1}\|\boldsymbol{\beta}\|^{-1} f_{\boldsymbol{u}^\top \boldsymbol{Z}}(\gamma)\boldsymbol{\nu}^\top \mathbb{E}\big(\boldsymbol{Z}\mathbb{1}\{\widehat{Y}^*(\boldsymbol{X}_1) = 1\}\big)\end{aligned}$$

and

$$\begin{aligned}\Lambda_{1,2} &= \Lambda_{2,1} \\ &= \Lambda_{1,1} + \|\boldsymbol{\beta}\|^{-1} f_{\boldsymbol{u}^\top \boldsymbol{Z}}(\gamma) \operatorname{Cov}\big(\boldsymbol{\beta}^\top \boldsymbol{Z}(\boldsymbol{\nu} - 2^{-1}\boldsymbol{\beta})^\top \boldsymbol{Z}, \mathbb{1}\{\widehat{Y}^*(\boldsymbol{X}_1) = 1\}\big) \\ &\quad + w_1^{-1}\|\boldsymbol{\beta}\|^{-1} f_{\boldsymbol{u}^\top \boldsymbol{Z}}(\gamma)\boldsymbol{\nu}^\top \mathbb{E}\big(\boldsymbol{Z}\mathbb{1}\{\widehat{Y}^*(\boldsymbol{X}_1) = 1\}\big).\end{aligned}$$

Instead of the standard estimator for the covariance matrix $\boldsymbol{\Sigma}$, one could use the more robust M-estimators defined in Section 1.1.2. Dümbgen et al. (2013) showed that these estimators satisfy the assumptions of Lemma 3.8 if $\mathcal{L}(\boldsymbol{Z})$ is spherically symmetric. In this case Condition (3.11) is not necessary and Condition (3.1) can be relaxed to

$$K(\boldsymbol{U}) := \int_{\mathbb{R}^{d-1}} \sup_{\boldsymbol{a} \in \boldsymbol{U}^\perp} \|\nabla f_{\boldsymbol{Z}}(\boldsymbol{a} + \boldsymbol{U}\boldsymbol{z})\| \, \mathrm{d}\boldsymbol{z} < \infty. \tag{3.12}$$

Theorem 3.10. *Let* $L = 2$ *and* $\widehat{\boldsymbol{\Sigma}}$ *be an estimator satisfying the assumptions of Lemma 3.8. Suppose that* $\mathcal{L}(\boldsymbol{Z})$ *is spherically symmetric satisfying* (3.12) *and* $\mathbb{E}(\|\boldsymbol{Z}\|^2) < \infty$. *Then for the standard linear classifier,*

$$\sqrt{n}\begin{pmatrix} R_1 - \mathbb{P}\big(\widehat{Y}_n^*(\boldsymbol{X}) \neq Y \;\big|\; Y = 1\big) \\ \widehat{R}_1 - \mathbb{P}\big(\widehat{Y}_n^*(\boldsymbol{X}) \neq Y \;\big|\; Y = 1\big) \end{pmatrix} \to_{\mathcal{L}} \mathcal{N}_2\big(\boldsymbol{0}, \boldsymbol{\Lambda}\big)$$

and the components of the covariance matrix $\boldsymbol{\Lambda}$ *are given by*

$$\begin{aligned}\Lambda_{1,1} = f_{Z_1}(\gamma)^2 \Big[& w_1(2^{-1}\|\boldsymbol{\beta}\| - \gamma)^2 \operatorname{Var}\Big(g_1\big(\|\boldsymbol{Z}\|\big) Z_1^2 - h_1\big(\|\boldsymbol{Z}\|\big)\Big) \\ & + w_2(2^{-1}\|\boldsymbol{\beta}\| - \gamma)^2 \operatorname{Var}\Big(g_2\big(\|\boldsymbol{Z}\|\big) Z_1^2 - h_2\big(\|\boldsymbol{Z}\|\big)\Big) \\ & + w_1^{-1}(\gamma/\|\boldsymbol{\beta}\|)^2 + w_2^{-1}(\gamma/\|\boldsymbol{\beta}\| - 1)^2 \\ & + \gamma(2\gamma/\|\boldsymbol{\beta}\| - 1)\mathbb{E}\Big(g_1\big(\|\boldsymbol{Z}\|\big) Z_1^3 - h_1\big(\|\boldsymbol{Z}\|\big) Z_1\Big) \\ & + (\|\boldsymbol{\beta}\| - \gamma)(2\gamma/\|\boldsymbol{\beta}\| - 1)\mathbb{E}\Big(g_2\big(\|\boldsymbol{Z}\|\big) Z_1^3 - h_2\big(\|\boldsymbol{Z}\|\big) Z_1\Big)\Big],\end{aligned}$$

$$\begin{aligned}\Lambda_{2,2} &= \Lambda_{1,1} + w_1^{-1}\big(\mathbb{P}(Z_1 \le \gamma) - \mathbb{P}(Z_1 \le \gamma)^2\big) \\ &\quad + f_{Z_1}(\gamma)(2\gamma - \|\boldsymbol{\beta}\|)\mathbb{E}\Big(\big(g_1(\|\boldsymbol{Z}\|)Z_1^2 - h_1(\|\boldsymbol{Z}\|)\big)\mathbb{1}\{Z_1 \le \gamma\}\Big) \\ &\quad + 2w_1^{-1} f_{Z_1}(\gamma)(\gamma/\|\boldsymbol{\beta}\|)\mathbb{E}\big(Z_1\mathbb{1}\{Z_1 \le \gamma\}\big)\end{aligned}$$

and

$$\begin{aligned}\Lambda_{1,2} &= \Lambda_{2,1} \\ &= \Lambda_{1,1} + f_{Z_1}(\gamma)(\gamma - \|\boldsymbol{\beta}\|/2)\mathbb{E}\Big(\big(g_1(\|\boldsymbol{Z}\|)Z_1^2 - h_1(\|\boldsymbol{Z}\|)\big)\mathbb{1}\{Z_1 \le \gamma\}\Big) \\ &\quad + w_1^{-1} f_{Z_1}(\gamma)(\gamma/\|\boldsymbol{\beta}\|)\mathbb{E}\big(Z_1\mathbb{1}\{Z_1 \le \gamma\}\big).\end{aligned}$$

Proof of Theorem 3.9. Note that

$$\begin{aligned}&\mathbb{P}\big(\widehat{Y}_n^*(\boldsymbol{X}) = 1 \,\big|\, Y = 1\big) \\ &\quad = \mathbb{P}\left((\boldsymbol{X} - \boldsymbol{\mu}_{1,2})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) + \log(\widehat{w}_2/\widehat{w}_1) \le 0 \,\middle|\, Y = 1\right) \\ &\quad = P_1\big\{\boldsymbol{x} \in \mathbb{R}^d \colon \boldsymbol{b}^\top \boldsymbol{x} + a_n \le 0\big\},\end{aligned}$$

where $a_n := (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{1,2} + \log(\widehat{w}_2/\widehat{w}_1)$, $\boldsymbol{b} := \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$ and

$$\begin{aligned}&\mathbb{P}\big(\widehat{Y}(\boldsymbol{X}, \mathcal{D}) = 1 \,\big|\, \mathcal{D}, Y = 1\big) \\ &\quad = \mathbb{P}\big((\boldsymbol{X} - \widehat{\boldsymbol{\mu}}_{\mathbf{1,2}})^\top \widehat{\boldsymbol{\Sigma}}^{-1}(\widehat{\boldsymbol{\mu}}_2 - \widehat{\boldsymbol{\mu}}_1) + \log(\widehat{w}_2/\widehat{w}_1) \le 0 \mid \mathcal{D}, Y = 1\big) \\ &\quad = P_1\big\{\boldsymbol{x} \in \mathbb{R}^d \colon \widehat{\boldsymbol{b}}^\top \boldsymbol{x} + \widehat{a} \le 0\big\},\end{aligned}$$

where $\widehat{a} := (\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2)^\top \widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}_{1,2} + \log(\widehat{w}_2/\widehat{w}_1)$ and $\widehat{\boldsymbol{b}} := \widehat{\boldsymbol{\Sigma}}^{-1}(\widehat{\boldsymbol{\mu}}_2 - \widehat{\boldsymbol{\mu}}_1)$ with estimators $\widehat{\boldsymbol{\mu}}_\theta(\mathcal{D})$ and $\widehat{\boldsymbol{\Sigma}}^{-1}(\mathcal{D})$. Elementary calculations reveal that $\widehat{\boldsymbol{\mu}}_\theta(\mathcal{D}_i(\boldsymbol{X})) = \widehat{\boldsymbol{\mu}}_\theta(\mathcal{D}) + \mathrm{O}_\mathrm{p}(n^{-1})$ and $\widehat{\boldsymbol{\Sigma}}^{-1}(\mathcal{D}_i(\boldsymbol{X})) = \widehat{\boldsymbol{\Sigma}}^{-1}(\mathcal{D}) + \mathrm{O}_\mathrm{p}(n^{-1})$. Thus

$$\begin{aligned}\widehat{R}_1 &= \frac{\#\{i \in \mathcal{G}_1 \colon \widehat{\boldsymbol{b}}(\mathcal{D}_i)^\top \boldsymbol{X}_i + \widehat{a}(\mathcal{D}_i) > 0\}}{N_1} \\ &= \frac{\#\{i \in \mathcal{G}_1 \colon \widehat{\boldsymbol{b}}(\mathcal{D})^\top \boldsymbol{X}_i + \widehat{a}(\mathcal{D}) + \mathrm{O}_\mathrm{p}(n^{-1}) > 0\}}{N_1} \\ &= \widehat{P}_1\big\{\boldsymbol{x} \in \mathbb{R}^d \colon \widehat{\boldsymbol{b}}(\mathcal{D})^\top \boldsymbol{X}_i + \widehat{a}' > 0\big\}\end{aligned}$$

with $\widehat{a}' = \widehat{a}(\mathcal{D}) + \mathrm{O}_\mathrm{p}(n^{-1})$.

Define $H_n := \{\boldsymbol{x} \in \mathbb{R}^d \colon \boldsymbol{b}^\top \boldsymbol{x} + a_n \le 0\}$, $\widehat{H} := \{\boldsymbol{x} \in \mathbb{R}^d \colon \widehat{\boldsymbol{b}}^\top \boldsymbol{x} + \widehat{a} \le 0\}$ and

$\widehat{H}' := \{\boldsymbol{x} \in \mathbb{R}^d \colon \widehat{\boldsymbol{b}}^\top \boldsymbol{x} + \widehat{a}' \le 0\}$. Then

$$\begin{aligned}\boldsymbol{\eta}_n &:= \sqrt{n}\begin{pmatrix} P_1(\widehat{H}) - P_1(H_n) \\ \widehat{P}_1(\widehat{H}') - P_1(H_n)\end{pmatrix} \\ &= -\sqrt{n}\begin{pmatrix} R_1 - \mathbb{P}\big(\widehat{Y}_n^*(\boldsymbol{X}) \neq Y \;\big|\; Y = 1\big) \\ \widehat{R}_1 - \mathbb{P}\big(\widehat{Y}_n^*(\boldsymbol{X}) \neq Y \;\big|\; Y = 1\big)\end{pmatrix}.\end{aligned}$$

Next note that

$$\begin{aligned}\widehat{\boldsymbol{b}} &= \widehat{\boldsymbol{\Sigma}}^{-1}(\widehat{\boldsymbol{\mu}}_2 - \widehat{\boldsymbol{\mu}}_1) \\ &= (\boldsymbol{\Sigma}^{-1} + n^{-1/2}\boldsymbol{\Delta}_{\boldsymbol{\Sigma}^{-1}})\big((\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) + n^{-1/2}(\boldsymbol{\Delta}_{\boldsymbol{\mu}_2} - \boldsymbol{\Delta}_{\boldsymbol{\mu}_1})\big) \\ &= \boldsymbol{b} + n^{-1/2}\big(\boldsymbol{\Delta}_{\boldsymbol{\Sigma}^{-1}}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) + \boldsymbol{\Sigma}^{-1}(\boldsymbol{\Delta}_{\boldsymbol{\mu}_2} - \boldsymbol{\Delta}_{\boldsymbol{\mu}_1})\big) + \mathrm{O_p}(n^{-1}).\end{aligned}$$

Hence

$$\begin{aligned}\boldsymbol{\Delta}_{\boldsymbol{b}} &:= \sqrt{n}(\widehat{\boldsymbol{b}} - \boldsymbol{b}) \\ &= \boldsymbol{\Delta}_{\boldsymbol{\Sigma}^{-1}}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) + \boldsymbol{\Sigma}^{-1}(\boldsymbol{\Delta}_{\boldsymbol{\mu}_2} - \boldsymbol{\Delta}_{\boldsymbol{\mu}_1}) + \mathrm{O_p}(n^{-1/2}) \\ &= \sum_{i=1}^n \boldsymbol{Y}_{n,i}^{\boldsymbol{b}} + \mathrm{o_p}(1),\end{aligned}$$

where $\boldsymbol{Y}_{n,i}^{\boldsymbol{b}} := \boldsymbol{Y}_{n,i}^{\boldsymbol{\Sigma}^{-1}}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) + \boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}_{n,i}^{\boldsymbol{\mu}_2} - \boldsymbol{Y}_{n,i}^{\boldsymbol{\mu}_1})$ with $\boldsymbol{Y}_{n,i}^{\boldsymbol{\Sigma}^{-1}}$ and $\boldsymbol{Y}_{n,i}^{\boldsymbol{\mu}_\theta}$ as in Lemma 3.6 and 3.7. Moreover,

$$\begin{aligned}\widehat{a} &= (\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2)^\top \widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}_{1,2} + \log(\widehat{w}_2/\widehat{w}_1) \\ &= a_n + n^{-1/2}\big((\boldsymbol{\Delta}_{\boldsymbol{\mu}_1} - \boldsymbol{\Delta}_{\boldsymbol{\mu}_2})^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{1,2} + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Delta}_{\boldsymbol{\Sigma}^{-1}}\boldsymbol{\mu}_{1,2} \\ &\qquad + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Delta}_{\boldsymbol{\mu}_{1,2}}\big) + \mathrm{O_p}(n^{-1})\end{aligned}$$

and

$$\begin{aligned}\Delta_a &:= \sqrt{n}(\widehat{a} - a_n) \\ &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Delta}_{\boldsymbol{\Sigma}^{-1}}\boldsymbol{\mu}_{1,2} + \boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Delta}_{\boldsymbol{\mu}_1} - \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Delta}_{\boldsymbol{\mu}_2} + \mathrm{O_p}(n^{-1/2}) \\ &= \sum_{i=1}^n Y_{n,i}^a + \mathrm{o_p}(1),\end{aligned}$$

where $Y_{n,i}^a := (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{Y}_{n,i}^{\boldsymbol{\Sigma}^{-1}}\boldsymbol{\mu}_{1,2} + \boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{Y}_{n,i}^{\boldsymbol{\mu}_1} - \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{Y}_{n,i}^{\boldsymbol{\mu}_2}$.

Define $\boldsymbol{\psi}_n^\top := (\boldsymbol{b}^\top, a_n)^\top$ and

$$\boldsymbol{\Delta}_{\boldsymbol{\psi}} := \sqrt{n}(\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}_n) = \begin{pmatrix}\boldsymbol{\Delta}_{\boldsymbol{b}} \\ \Delta_a\end{pmatrix} = \sum_{i=1}^n \boldsymbol{Y}_{n,i}^{\boldsymbol{\psi}} + \mathrm{o_p}(1)$$

with

$$\boldsymbol{Y}_{n,i}^{\boldsymbol{\psi}} := \begin{pmatrix} \boldsymbol{Y}_{n,i}^{\boldsymbol{b}} \\ Y_{n,i}^{a} \end{pmatrix}.$$

By Lemma 3.1,

$$\begin{aligned}\sqrt{n}\big(P_1(\widehat{H}) - P_1(H_n)\big) &= \sqrt{n}\big(P_1(\widehat{H}') - P_1(H_n)\big) + \mathrm{o_p}(1) \\ &= \boldsymbol{c}^\top \boldsymbol{\Delta}_{\boldsymbol{\psi}} + \mathrm{o_p}(1) \\ &= \sum_{i=1}^{n} \boldsymbol{c}^\top \boldsymbol{Y}_{n,i}^{\boldsymbol{\psi}} + \mathrm{o_p}(1),\end{aligned}$$

where

$$\begin{aligned}\boldsymbol{c} = \boldsymbol{c}_n &= -f_{\boldsymbol{b}^\top \boldsymbol{X}_1}(-a_n)\mathbb{E}\Big(\begin{pmatrix} \boldsymbol{X}_1 \\ 1 \end{pmatrix} \Big| \boldsymbol{b}^\top \boldsymbol{X}_1 = -a_n \Big)^\top \\ &= -f_{\boldsymbol{b}^\top \boldsymbol{X}_1}(-a_n) \begin{pmatrix} \boldsymbol{\Sigma}^{1/2}\boldsymbol{\nu}_n + \boldsymbol{\mu}_1 \\ 1 \end{pmatrix}\end{aligned}$$

and $\boldsymbol{\nu}_n := \mathbb{E}(\boldsymbol{Z} \mid \boldsymbol{b}^\top \boldsymbol{X}_1 = -a_n)$.

Regarding the second component of $\boldsymbol{\eta}_n$, note that

$$\widehat{P}_1(\widehat{H}') - P_1(H_n) = \big(P_1(\widehat{H}') - P_1(H_n)\big) + \big(\widehat{P}_1 - P_1\big)(H_n) + R_n$$

with

$$\begin{aligned}R_n &:= \big(\widehat{P}_1 - P_1\big)(\widehat{H}') - \big(\widehat{P}_1 - P_1\big)(H_n) \\ &= n^{-1/2}\big(\mathbb{B}_{P_1,n}(\widehat{\boldsymbol{b}}, \widehat{a}') - \mathbb{B}_{P_1,n}(\boldsymbol{b}, a_n)\big)\end{aligned}$$

and the empirical process $\mathbb{B}_{P_1,n}$ defined in Section 3.1.2. Lemma 3.1 entails that $P_1(\widehat{H}' \triangle H_n) \to_p 0$ and we deduce from Theorem 3.5 that $R_n = \mathrm{o_p}(n^{-1/2})$. Next we define

$$Y_{n,i}^{p} := \frac{\mathbb{1}\{i \in \mathcal{G}_1\}}{\sqrt{n}\widehat{w}_1}\big(\mathbb{1}\{\boldsymbol{X}_i \in H_n\} - P_1(H_n)\big)$$

such that

$$\begin{aligned}\sum_{i=1}^{n} Y_{n,i}^{p} &= \frac{\sqrt{n}}{N_1} \sum_{i \in \mathcal{G}_1} \big(\mathbb{1}\{\boldsymbol{X}_i \in H_n\} - P_1(H_n)\big) \\ &= \sqrt{n}\big(\widehat{P}_1 - P_1\big)(H_n)\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{\eta}_n &= \sqrt{n}\begin{pmatrix} P_1(\widehat{H}) - P_1(H_n) \\ \big(P_1(\widehat{H}') - P_1(H_n)\big) + \big(\widehat{P}_1 - P_1\big)(H_n)\end{pmatrix} + \mathrm{o}_{\mathrm{p}}(1) \\ &= \sum_{i=1}^{n} \boldsymbol{Y}_{n,i}^{\boldsymbol{\eta}} + \mathrm{o}_{\mathrm{p}}(1)\end{aligned}$$

with

$$\boldsymbol{Y}_{n,i}^{\boldsymbol{\eta}} := \begin{pmatrix} \boldsymbol{c}^\top \boldsymbol{Y}_{n,i}^{\boldsymbol{\psi}} \\ \boldsymbol{c}^\top \boldsymbol{Y}_{n,i}^{\boldsymbol{\psi}} + Y_{n,i}^{p}\end{pmatrix}.$$

Before we can apply the central limit theorem to $\sum_{i=1}^n \boldsymbol{Y}_{n,i}^{\boldsymbol{\eta}}$ we have to compute $\operatorname{Var}(\boldsymbol{Y}_{n,i}^{\boldsymbol{\eta}})$, the sum $\boldsymbol{\Lambda}^n := \sum_{i=1}^n \operatorname{Var}(\boldsymbol{Y}_{n,i}^{\boldsymbol{\eta}})$ and its limit $\boldsymbol{\Lambda} = \lim_{n\to\infty} \boldsymbol{\Lambda}^n$. To this end note that

$$\begin{aligned}\boldsymbol{c}^\top \boldsymbol{Y}_{n,i}^{\boldsymbol{\psi}} &= \boldsymbol{c}_{1:d}^\top \boldsymbol{Y}_{n,i}^{\boldsymbol{b}} + c_{d+1} Y_{n,i}^{a} \\ &= \boldsymbol{c}_{1:d}^\top \boldsymbol{Y}_{n,i}^{\boldsymbol{\Sigma}^{-1}}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) + \boldsymbol{c}_{1:d}^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}_{n,i}^{\boldsymbol{\mu}_2} - \boldsymbol{Y}_{n,i}^{\boldsymbol{\mu}_1}) \\ &\quad + c_{d+1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{Y}_{n,i}^{\boldsymbol{\Sigma}^{-1}} \boldsymbol{\mu}_{1,2} + c_{d+1} \boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{Y}_{n,i}^{\boldsymbol{\mu}_1} \\ &\quad - c_{d+1} \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{Y}_{n,i}^{\boldsymbol{\mu}_2} \\ &= (\boldsymbol{c}_{1:d} - c_{d+1}\boldsymbol{\mu}_{1,2})^\top \boldsymbol{Y}_{n,i}^{\boldsymbol{\Sigma}^{-1}}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \\ &\quad + (\boldsymbol{c}_{1:d} - c_{d+1}\boldsymbol{\mu}_{Y_i})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}_{n,i}^{\boldsymbol{\mu}_2} - \boldsymbol{Y}_{n,i}^{\boldsymbol{\mu}_1}) \\ &= \big((\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \otimes (\boldsymbol{c}_{1:d} - c_{d+1}\boldsymbol{\mu}_{1,2})\big)^\top \operatorname{vec}(\boldsymbol{Y}_{n,i}^{\boldsymbol{\Sigma}^{-1}}) \\ &\quad + \frac{\sqrt{n}}{N_{Y_i}} (-1)^{\mathbb{1}\{Y_i=1\}} (\boldsymbol{c}_{1:d} - c_{d+1}\boldsymbol{\mu}_{Y_i})^\top \boldsymbol{\Sigma}^{-1} \tilde{\boldsymbol{X}}_i \\ &= -\boldsymbol{v}_n^\top \operatorname{vec}(\boldsymbol{Y}_{n,i}^{\boldsymbol{\Sigma}}) + \frac{\sqrt{n}}{N_{Y_i}} (-1)^{\mathbb{1}\{Y_i=1\}} (\boldsymbol{c}_{1:d} - c_{d+1}\boldsymbol{\mu}_{Y_i})^\top \boldsymbol{\Sigma}^{-1} \tilde{\boldsymbol{X}}_i,\end{aligned}$$

where

$$\begin{aligned}\boldsymbol{v}_n &:= \big(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\big) \otimes \big(\boldsymbol{\Sigma}^{-1}(\boldsymbol{c}_{1:d} - c_{d+1}\boldsymbol{\mu}_{1,2})\big) \in \mathbb{R}^{d^2} \\ &= -f_{\boldsymbol{b}^\top \boldsymbol{X}_1}(-a_n)\boldsymbol{b} \otimes \big(\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\nu}_n - 2^{-1}\boldsymbol{\beta})\big).\end{aligned}$$

Here and for the following computations we utilize several times the relations

(1.22), (1.23) and (3.10). The upper left component of $\mathbf{\Lambda}$ is given by

$$\begin{aligned}
\Lambda_{1,1}^n &= \sum_{i=1}^n \operatorname{Var}(\boldsymbol{c}^\top \boldsymbol{Y}_{n,i}^{\boldsymbol{\psi}}) \\
&= \sum_{i=1}^n \Big(\operatorname{Var}\left(\boldsymbol{v}_n^\top \operatorname{vec}(\boldsymbol{Y}_{n,i}^{\boldsymbol{\Sigma}})\right) \\
&\qquad + \frac{n}{N_{Y_i}^2}(\boldsymbol{c}_{1:d} - c_{d+1}\boldsymbol{\mu}_{Y_i})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{c}_{1:d} - c_{d+1}\boldsymbol{\mu}_{Y_i}) \\
&\qquad + \frac{2\sqrt{n}(-1)^{\mathbb{1}\{Y_i=2\}}}{N_{Y_i}} \boldsymbol{v}_n^\top \operatorname{Cov}\left(\operatorname{vec}(\boldsymbol{Y}_{n,i}^{\boldsymbol{\Sigma}}), \tilde{\boldsymbol{X}}_i\right) \boldsymbol{\Sigma}^{-1}(\boldsymbol{c}_{1:d} - c_{d+1}\boldsymbol{\mu}_{Y_i}) \Big) \\
&= \operatorname{Var}\left(\boldsymbol{v}_n^\top \operatorname{vec}(\tilde{\boldsymbol{X}}_1 \tilde{\boldsymbol{X}}_1^\top)\right) + f_{\boldsymbol{b}^\top \boldsymbol{X}_1}(-a_n)^2 \big(\widehat{w}_1^{-1} \|\boldsymbol{\nu}_n\|^2 + \widehat{w}_2^{-1} \|\boldsymbol{\nu}_n - \boldsymbol{\beta}\|^2 \big) \\
&\qquad + 2c_{d+1} \boldsymbol{v}_n^\top \operatorname{Cov}\left(\operatorname{vec}(\tilde{\boldsymbol{X}}_1 \tilde{\boldsymbol{X}}_1^\top), \tilde{\boldsymbol{X}}_1\right) \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \\
&\to \|\boldsymbol{\beta}\|^{-2} f_{\boldsymbol{u}^\top \boldsymbol{Z}}(\gamma)^2 \Big[\operatorname{Var}\left(\boldsymbol{\beta}^\top \boldsymbol{Z} (\boldsymbol{\nu} - 2^{-1}\boldsymbol{\beta})^\top \boldsymbol{Z}\right) + w_1^{-1} \|\boldsymbol{\nu}\|^2 \\
&\qquad\qquad + w_2^{-1} \|\boldsymbol{\nu} - \boldsymbol{\beta}\|^2 + 2\mathbb{E}\big((\boldsymbol{\beta}^\top \boldsymbol{Z})^2 (\boldsymbol{\nu} - 2^{-1}\boldsymbol{\beta})^\top \boldsymbol{Z} \big) \Big],
\end{aligned}$$

where $\boldsymbol{u} := \|\boldsymbol{\beta}\|^{-1}\boldsymbol{\beta}$. We used that $f_{\boldsymbol{b}^\top \boldsymbol{X}_1}(-a) = \|\boldsymbol{\beta}\|^{-1} f_{\boldsymbol{u}^\top \boldsymbol{Z}}(\gamma)$ and $\boldsymbol{\nu}_n \to \boldsymbol{\nu}$. The latter assertion can be derived from (3.1).

To compute $\operatorname{Var}(\boldsymbol{c}^\top \boldsymbol{Y}_{n,i}^{\boldsymbol{\psi}} + Y_{n,i}^p)$ we need

$$\operatorname{Var}(Y_{n,i}^p) = \frac{\mathbb{1}\{i \in \mathcal{G}_1\}}{n \widehat{w}_1^2} \big(P_1(H_n) - P_1(H_n)^2 \big)$$

and

$$\begin{aligned}
&\operatorname{Cov}(\boldsymbol{c}^\top \boldsymbol{Y}_{n,i}^{\boldsymbol{\psi}}, Y_{n,i}^p) \\
&\quad = \operatorname{Cov}\left(- \boldsymbol{v}_n^\top \operatorname{vec}(\boldsymbol{Y}_{n,i}^{\boldsymbol{\Sigma}}), Y_{n,i}^p \right) \\
&\qquad + \operatorname{Cov}\left(\frac{\sqrt{n}}{N_{Y_i}} (-1)^{\mathbb{1}\{Y_i=1\}} (\boldsymbol{c}_{1:d} - c_{d+1}\boldsymbol{\mu}_{Y_i})^\top \boldsymbol{\Sigma}^{-1} \tilde{\boldsymbol{X}}_i, Y_{n,i}^p \right) \\
&\quad = - \frac{\mathbb{1}\{i \in \mathcal{G}_1\}}{N_1} \boldsymbol{v}_n^\top \operatorname{Cov}\left(\operatorname{vec}(\tilde{\boldsymbol{X}}_1 \tilde{\boldsymbol{X}}_1^\top), \mathbb{1}\{\boldsymbol{X}_1 \in H_n\} \right) \\
&\qquad - \frac{\mathbb{1}\{i \in \mathcal{G}_1\}}{N_1 \widehat{w}_1} (\boldsymbol{c}_{1:d} - c_{d+1}\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1} \operatorname{Cov}\left(\tilde{\boldsymbol{X}}_1, \mathbb{1}\{\boldsymbol{X}_1 \in H_n\} \right)
\end{aligned}$$

$$\begin{aligned}
&= \frac{\mathbb{1}\{i \in \mathcal{G}_1\}}{N_1} f_{\boldsymbol{b}^\top \boldsymbol{X}_1}(-a_n) \operatorname{Cov}\left(\boldsymbol{\beta}^\top \boldsymbol{Z} (\boldsymbol{\nu}_n - 2^{-1}\boldsymbol{\beta})^\top \boldsymbol{Z}, \mathbb{1}\{\boldsymbol{X}_1 \in H_n\} \right) \\
&\quad + \frac{\mathbb{1}\{i \in \mathcal{G}_1\}}{N_1 \widehat{w}_1} f_{\boldsymbol{b}^\top \boldsymbol{X}_1}(-a_n) \boldsymbol{\nu}_n^\top \mathbb{E}\big(\boldsymbol{Z} \mathbb{1}\{\boldsymbol{X}_1 \in H_n\} \big).
\end{aligned}$$

Employing (3.1) again, one can show that $P_1(H_n) \to \mathbb{P}(\widehat{Y}^*(\boldsymbol{X}_1) = 1)$. Thus $\operatorname{Cov}\big(\boldsymbol{Z}, \mathbb{1}\{\boldsymbol{X}_1 \in H\}\big) \to \operatorname{Cov}\big(\boldsymbol{Z}, \mathbb{1}\{\widehat{Y}^*(\boldsymbol{X}_1) = 1\}\big)$ and $\operatorname{Cov}\big(\boldsymbol{\beta}^\top \boldsymbol{Z}(\boldsymbol{\nu}_n - 2^{-1}\boldsymbol{\beta})^\top \boldsymbol{Z}, \mathbb{1}\{\boldsymbol{X}_1 \in H\}\big) \to \operatorname{Cov}\big(\boldsymbol{\beta}^\top \boldsymbol{Z}(\boldsymbol{\nu} - 2^{-1}\boldsymbol{\beta})^\top \boldsymbol{Z}, \mathbb{1}\{\widehat{Y}^*(\boldsymbol{X}_1) = 1\}\big)$ by dominated convergence. Hence the lower left component of $\boldsymbol{\Lambda}$ is given by

$$\begin{aligned}
\Lambda_{2,2}^n &= \sum_{i=1}^n \operatorname{Var}(\boldsymbol{c}^\top \boldsymbol{Y}_{n,i}^{\boldsymbol{\psi}} + Y_{n,i}^p) \\
&\to \Lambda_{1,1} + w_1^{-1}\big(\mathbb{P}(\widehat{Y}^*(\boldsymbol{X}_1) = 1) - \mathbb{P}(\widehat{Y}^*(\boldsymbol{X}_1) = 1)^2\big) \\
&\quad + 2\|\boldsymbol{\beta}\|^{-1} f_{\boldsymbol{u}^\top \boldsymbol{Z}}(\gamma) \operatorname{Cov}\big(\boldsymbol{\beta}^\top \boldsymbol{Z}(\boldsymbol{\nu} - 2^{-1}\boldsymbol{\beta})^\top \boldsymbol{Z}, \mathbb{1}\{\widehat{Y}^*(\boldsymbol{X}_1) = 1\}\big) \\
&\quad + 2w_1^{-1}\|\boldsymbol{\beta}\|^{-1} f_{\boldsymbol{u}^\top \boldsymbol{Z}}(\gamma) \boldsymbol{\nu}^\top \mathbb{E}\big(\boldsymbol{Z}\mathbb{1}\{\widehat{Y}^*(\boldsymbol{X}_1) = 1\}\big)
\end{aligned}$$

and the other components are equal to

$$\begin{aligned}
\Lambda_{1,2}^n &= \Lambda_{2,1}^n \\
&= \sum_{i=1}^n \big(\operatorname{Var}(\boldsymbol{c}^\top \boldsymbol{Y}_{n,i}^{\boldsymbol{\psi}}) + \operatorname{Cov}(\boldsymbol{c}^\top \boldsymbol{Y}_{n,i}^{\boldsymbol{\psi}}, Y_{n,i}^p)\big) \\
&\to \Lambda_{1,1} + \|\boldsymbol{\beta}\|^{-1} f_{\boldsymbol{u}^\top \boldsymbol{Z}}(\gamma) \operatorname{Cov}\big(\boldsymbol{\beta}^\top \boldsymbol{Z}(\boldsymbol{\nu} - 2^{-1}\boldsymbol{\beta})^\top \boldsymbol{Z}, \mathbb{1}\{\widehat{Y}^*(\boldsymbol{X}_1) = 1\}\big) \\
&\quad + w_1^{-1}\|\boldsymbol{\beta}\|^{-1} f_{\boldsymbol{u}^\top \boldsymbol{Z}}(\gamma) \boldsymbol{\nu}^\top \mathbb{E}\big(\boldsymbol{Z}\mathbb{1}\{\widehat{Y}^*(\boldsymbol{X}_1) = 1\}\big).
\end{aligned}$$

Next we show that Lindeberg's condition is satisfied. Note that

$$\begin{aligned}
\sum_{i=1}^n \mathbb{E}\left(\|\boldsymbol{Y}_{n,i}^{\boldsymbol{\eta}}\|^2 \min(1, \|\boldsymbol{Y}_{n,i}^{\boldsymbol{\eta}}\|)\right) &= \sum_{\lambda \in \{1,2\}} \sum_{i \in \mathcal{G}_\lambda} \mathbb{E}\left(\|\boldsymbol{Y}_{n,i}^{\boldsymbol{\eta}}\|^2 \min(1, \|\boldsymbol{Y}_{n,i}^{\boldsymbol{\eta}}\|)\right) \\
&= \sum_{\lambda \in \{1,2\}} N_\lambda \mathbb{E}\left(\|\boldsymbol{Y}_{n,\lambda}^{\boldsymbol{\eta}}\|^2 \min(1, \|\boldsymbol{Y}_{n,\lambda}^{\boldsymbol{\eta}}\|)\right)
\end{aligned}$$

and for $\lambda \in \{1,2\}$, $\|\boldsymbol{Y}_{n,\lambda}^{\boldsymbol{\eta}}\| \le 2\|\boldsymbol{c}\|\|\boldsymbol{Y}_{n,\lambda}^{\boldsymbol{\psi}}\| + |Y_{n,\lambda}^p| \le 2\|\boldsymbol{c}\|\|\boldsymbol{Y}_{n,\lambda}^{\boldsymbol{\psi}}\| + n^{-1/2}\widehat{w}_1^{-1}$. For some constants c' and c''

$$\|\boldsymbol{Y}_{n,\lambda}^{\boldsymbol{\psi}}\| \le \|\boldsymbol{Y}_{n,\lambda}^{\boldsymbol{b}}\| + |Y_{n,\lambda}^a| \le c'\|\boldsymbol{Y}_{n,\lambda}^{\boldsymbol{\Sigma}}\| + c''\frac{\sqrt{n}}{N_\lambda}\|\tilde{\boldsymbol{X}}_\lambda\|.$$

Therefore $\|\boldsymbol{Y}_{n,\lambda}^{\boldsymbol{\Sigma}}\| \to_{\text{a.s.}} 0$ implies $\|\boldsymbol{Y}_{n,\lambda}^{\boldsymbol{\psi}}\| \to_{\text{a.s.}} 0$ and $\|\boldsymbol{Y}_{n,\lambda}^{\boldsymbol{\eta}}\| \to_{\text{a.s.}} 0$. Since

$$\mathbb{E}\|\tilde{\boldsymbol{X}}_\lambda\|^2 < \infty, \quad \mathbb{E}\sqrt{n}\|\boldsymbol{Y}_{n,\lambda}^{\boldsymbol{\Sigma}}\|\|\tilde{\boldsymbol{X}}_\lambda\| < \infty \quad \textit{and} \quad \mathbb{E}n\|\boldsymbol{Y}_{n,\lambda}^{\boldsymbol{\Sigma}}\|^2 < \infty$$

by assumption, the dominated convergence theorem implies that

$$N_\lambda \mathbb{E}\left(\|\boldsymbol{Y}_{n,\lambda}^{\boldsymbol{\eta}}\|^2 \min(1, \|\boldsymbol{Y}_{n,\lambda}^{\boldsymbol{\eta}}\|)\right) \to 0$$

for any $\lambda \in \mathcal{Y}$ and thus Lindeberg's condition

$$\sum_{i=1}^{n} \mathbb{E}\left(\|\boldsymbol{Y}_{n,i}^{\boldsymbol{\eta}}\|^2 \min(1, \|\boldsymbol{Y}_{n,i}^{\boldsymbol{\eta}}\|)\right) \to 0$$

is satisfied. Since $\mathbb{E}(\boldsymbol{Y}_{n,i}^{\boldsymbol{\eta}}) = \mathbf{0}$ for all $i \leq n$ and the Gaussian distribution is symmetric, the assertion follows from the central limit theorem (Theorem A.1). □

Proof of Theorem 3.10. The prof of this theorem is similar to the proof of Theorem 3.9. But the covariance matrix of the limit distribution is slightly different. The elliptical symmetry implies that

$$\begin{aligned}
&\mathbb{P}\big(\widehat{Y}^*(\boldsymbol{X}) = 1 \,\big|\, Y = 1\big) \\
&\quad = \mathbb{P}\big((\boldsymbol{X} - \boldsymbol{\mu}_{1,2})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) + \log(\widehat{w}_2/\widehat{w}_1) \leq 0 \mid Y = 1\big) \\
&\quad = P_1\{\boldsymbol{x} \in \mathbb{R}^d \colon \boldsymbol{b}^\top \boldsymbol{x} + a_n \leq 0\} \\
&\quad = P_1\{\boldsymbol{\beta}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}_1) + \alpha_n \leq 0\} \\
&\quad = P_0\{\boldsymbol{\beta}^\top \boldsymbol{x} + \alpha_n \leq 0\} \\
&\quad = P_0\{(\boldsymbol{\beta}/\|\boldsymbol{\beta}\|)^\top \boldsymbol{x} + \alpha_n/\|\boldsymbol{\beta}\| \leq 0\} \\
&\quad = P_0\{x_1 + \alpha_n/\|\boldsymbol{\beta}\| \leq 0\},
\end{aligned}$$

where $\alpha_n = a_n + \boldsymbol{b}^\top \boldsymbol{\mu}_1$. Analogously we get

$$\mathbb{P}\big(\widehat{Y}(\boldsymbol{X}, \mathcal{D}) = 1 \,\big|\, \mathcal{D}, Y = 1\big) = P_0\{x_1 + \widehat{\alpha}/\|\widehat{\boldsymbol{\beta}}\| \leq 0\}$$

with $\widehat{\boldsymbol{\beta}} = \boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{b}}$ and $\widehat{\alpha} = \widehat{a} + \widehat{\boldsymbol{b}}^\top \boldsymbol{\mu}_1$.

Employing Lemma 3.1 with $\widehat{\boldsymbol{b}} = \boldsymbol{b} = \boldsymbol{e}_1$, $a = \alpha_n/\|\boldsymbol{\beta}\|$ and $\widehat{a} = \widehat{\alpha}/\|\widehat{\boldsymbol{\beta}}\| = \alpha_n/\|\boldsymbol{\beta}\| + \mathrm{O_p}(n^{-1/2})$ yields

$$\begin{aligned}
\sqrt{n}\big(P_1(\widehat{H}) - P_1(H_n)\big) &= \boldsymbol{c}^\top \boldsymbol{\Delta}_{\boldsymbol{\psi}} + \mathrm{o_p}(1) \\
&= -f_{Z_1}(\gamma_n)\sqrt{n}\Big(\frac{\widehat{\alpha}}{\|\widehat{\boldsymbol{\beta}}\|} - \frac{\alpha_n}{\|\boldsymbol{\beta}\|}\Big) + \mathrm{o_p}(1),
\end{aligned}$$

where $\gamma_n := \|\boldsymbol{\beta}\|/2 + \log(\widehat{w}_1/\widehat{w}_2)/\|\boldsymbol{\beta}\| = -\alpha_n/\|\boldsymbol{\beta}\|$. Note that we use Lemma 3.1 with $\widehat{\boldsymbol{b}} = \boldsymbol{b} = \boldsymbol{e}_1$. Therefore Condition (3.2) is not necessary and Condition (3.3) can be relaxed to (3.12).

The first order Taylor expansion of $\|\widehat{\boldsymbol{\beta}}\|$ is given by

$$\|\widehat{\boldsymbol{\beta}}\| = \|\boldsymbol{\beta}\| + \frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \boldsymbol{\beta}}{\|\boldsymbol{\beta}\|} + \mathrm{o_p}(n^{-1/2}).$$

Together with $\widehat{\alpha} = \alpha_n + \mathrm{O_p}(n^{-1/2})$ this entails that

$$\begin{aligned}\frac{\widehat{\alpha}}{\|\widehat{\boldsymbol{\beta}}\|} - \frac{\widehat{\alpha}}{\|\boldsymbol{\beta}\|} &= \frac{-(\|\widehat{\boldsymbol{\beta}}\| - \|\boldsymbol{\beta}\|)\alpha_n}{(\|\widehat{\boldsymbol{\beta}}\| - \|\boldsymbol{\beta}\|)\|\boldsymbol{\beta}\| + \|\boldsymbol{\beta}\|^2} + \mathrm{o_p}(n^{-1/2}) \\ &= \frac{-(\|\widehat{\boldsymbol{\beta}}\| - \|\boldsymbol{\beta}\|)\alpha_n}{\|\boldsymbol{\beta}\|^2} + \mathrm{o_p}(n^{-1/2}) \\ &= -(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \boldsymbol{\beta} \frac{\alpha_n}{\|\boldsymbol{\beta}\|^3} + \mathrm{o_p}(n^{-1/2})\end{aligned}$$

and thus

$$\begin{aligned}&\sqrt{n}\Big(\frac{\widehat{\alpha}}{\|\widehat{\boldsymbol{\beta}}\|} - \frac{\alpha_n}{\|\boldsymbol{\beta}\|}\Big) \\ &\quad= -\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \boldsymbol{\beta} \frac{\alpha_n}{\|\boldsymbol{\beta}\|^3} + \sqrt{n}\frac{\widehat{\alpha} - \alpha_n}{\|\boldsymbol{\beta}\|} + \mathrm{o_p}(1) \\ &\quad= -\frac{\alpha_n}{\|\boldsymbol{\beta}\|^3}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\top \boldsymbol{\Delta_b} + \frac{1}{\|\boldsymbol{\beta}\|}\boldsymbol{\Delta}_a + \frac{1}{\|\boldsymbol{\beta}\|}\boldsymbol{\mu}_1^\top \boldsymbol{\Delta_b} + \mathrm{o_p}(1) \\ &\quad= \Big(\Big(\frac{\alpha_n}{\|\boldsymbol{\beta}\|^3} + \frac{1}{\|\boldsymbol{\beta}\|}\Big)\boldsymbol{\mu}_1 - \frac{\alpha_n}{\|\boldsymbol{\beta}\|^3}\boldsymbol{\mu}_2\Big)^\top \boldsymbol{\Delta_b} + \frac{1}{\|\boldsymbol{\beta}\|}\boldsymbol{\Delta}_a + \mathrm{o_p}(1).\end{aligned}$$

Therefore in the elliptic symmetric case $\boldsymbol{c}$ is given by

$$\begin{aligned}\boldsymbol{c} &= -\frac{f_{Z_1}(\gamma_n)}{\|\boldsymbol{\beta}\|}\begin{pmatrix}(\frac{\alpha_n}{\|\boldsymbol{\beta}\|^2} + 1)\boldsymbol{\mu}_1 - \frac{\alpha_n}{\|\boldsymbol{\beta}\|^2}\boldsymbol{\mu}_2 \\ 1\end{pmatrix} \\ &= -\frac{f_{Z_1}(\gamma_n)}{\|\boldsymbol{\beta}\|}\begin{pmatrix}(1 - \frac{\gamma_n}{\|\boldsymbol{\beta}\|})\boldsymbol{\mu}_1 + \frac{\gamma_n}{\|\boldsymbol{\beta}\|}\boldsymbol{\mu}_2 \\ 1\end{pmatrix}\end{aligned}$$

and

$$\boldsymbol{v}_n = \frac{f_{Z_1}(\gamma_n)}{\|\boldsymbol{\beta}\|}\Big(\frac{1}{2} - \frac{\gamma_n}{\|\boldsymbol{\beta}\|}\Big)(\boldsymbol{b} \otimes \boldsymbol{b}).$$

Hence the upper left component of $\boldsymbol{\Lambda}$ is given by

$$\begin{aligned}\Lambda_{1,1} = f_{Z_1}(\gamma)^2\Big[&w_1(2^{-1}\|\boldsymbol{\beta}\| - \gamma)^2 \operatorname{Var}\Big(g_1\big(\|\boldsymbol{Z}\|\big)Z_1^2 - h_1\big(\|\boldsymbol{Z}\|\big)\Big) \\ &+ w_2(2^{-1}\|\boldsymbol{\beta}\| - \gamma)^2 \operatorname{Var}\Big(g_2\big(\|\boldsymbol{Z}\|\big)Z_1^2 - h_2\big(\|\boldsymbol{Z}\|\big)\Big) \\ &+ w_1^{-1}(\gamma/\|\boldsymbol{\beta}\|)^2 + w_2^{-1}(\gamma/\|\boldsymbol{\beta}\| - 1)^2 \\ &+ \gamma(2\gamma/\|\boldsymbol{\beta}\| - 1)\mathbb{E}\Big(g_1\big(\|\boldsymbol{Z}\|\big)Z_1^3 - h_1\big(\|\boldsymbol{Z}\|\big)Z_1\Big) \\ &+ (\|\boldsymbol{\beta}\| - \gamma)(2\gamma/\|\boldsymbol{\beta}\| - 1)\mathbb{E}\Big(g_2\big(\|\boldsymbol{Z}\|\big)Z_1^3 - h_2\big(\|\boldsymbol{Z}\|\big)Z_1\Big)\Big].\end{aligned}$$

We utilized that $\|\boldsymbol{\beta}\|^{-1}\boldsymbol{\beta}^\top \boldsymbol{Z}$ has the same distribution as Z_1. Note that $P_1(H_n) = \mathbb{P}(Z_1 \leq \gamma_n)$, $\mathbb{1}\{\boldsymbol{X}_1 \in H_n\} = \mathbb{1}\{\|\boldsymbol{\beta}\|^{-1}\boldsymbol{\beta}^\top \boldsymbol{Z} \leq \gamma_n\}$ and

$$
\begin{aligned}
&\operatorname{Cov}(\boldsymbol{c}^\top \boldsymbol{Y}_{n,i}^{\boldsymbol{\psi}}, Y_{n,i}^{p}) \\
&= -\frac{\mathbb{1}\{i \in \mathcal{G}_1\}}{N_1} \boldsymbol{v}_n^\top \operatorname{Cov}\left(\sqrt{n}\operatorname{vec}(\boldsymbol{Y}_{n,1}^{\boldsymbol{\Sigma}}), \mathbb{1}\{\boldsymbol{X}_1 \in H_n\}\right) \\
&\quad - \frac{\mathbb{1}\{i \in \mathcal{G}_1\}}{N_1 \widehat{w}_1} (\boldsymbol{c}_{1:d} - c_{d+1}\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1} \operatorname{Cov}\left(\tilde{\boldsymbol{X}}_1, \mathbb{1}\{\boldsymbol{X}_1 \in H_n\}\right) \\
&= \frac{\mathbb{1}\{i \in \mathcal{G}_1\}}{N_1} f_{Z_1}(\gamma_n)(\gamma_n - \|\boldsymbol{\beta}\|/2)\mathbb{E}\Big(\big(g_1(\|\boldsymbol{Z}\|)Z_1^2 - h_1(\|\boldsymbol{Z}\|)\big)\mathbb{1}\{\boldsymbol{X}_1 \in H_n\}\Big) \\
&\quad + \frac{\mathbb{1}\{i \in \mathcal{G}_1\}}{N_1 \widehat{w}_1} \frac{\gamma_n f_{Z_1}(\gamma_n)}{\|\boldsymbol{\beta}\|} \mathbb{E}\big(Z_1 \mathbb{1}\{\boldsymbol{X}_1 \in H_n\}\big) \\
&= \frac{\mathbb{1}\{i \in \mathcal{G}_1\}}{N_1} f_{Z_1}(\gamma_n)(\gamma_n - \|\boldsymbol{\beta}\|/2)\mathbb{E}\Big(\big(g_1(\|\boldsymbol{Z}\|)Z_1^2 - h_1(\|\boldsymbol{Z}\|)\big)\mathbb{1}\{Z_1 \leq \gamma_n\}\Big) \\
&\quad + \frac{\mathbb{1}\{i \in \mathcal{G}_1\}}{N_1 \widehat{w}_1} \frac{\gamma_n f_{Z_1}(\gamma_n)}{\|\boldsymbol{\beta}\|} \mathbb{E}\big(Z_1 \mathbb{1}\{Z_1 \leq \gamma_n\}\big).
\end{aligned}
$$

Therefore the remaining components are given by

$$
\begin{aligned}
\Lambda_{2,2}^n &= \Lambda_{1,1}^n + \widehat{w}_1^{-1}\big(P_1(H_n) - P_1(H_n)^2\big) \\
&\quad + \frac{\mathbb{1}\{i \in \mathcal{G}_1\}}{N_1} f_{Z_1}(\gamma_n)(2\gamma_n - \|\boldsymbol{\beta}\|) \\
&\qquad\qquad \cdot \mathbb{E}\Big(\big(g_1(\|\boldsymbol{Z}\|)Z_1^2 - h_1(\|\boldsymbol{Z}\|)\big)\mathbb{1}\{Z_1 \leq \gamma_n\}\Big) \\
&\quad + 2\frac{\mathbb{1}\{i \in \mathcal{G}_1\}}{N_1 \widehat{w}_1} \frac{\gamma_n f_{Z_1}(\gamma_n)}{\|\boldsymbol{\beta}\|} \mathbb{E}\big(Z_1 \mathbb{1}\{Z_1 \leq \gamma_n\}\big). \\
&\to \Lambda_{1,1} + w_1^{-1}\big(\mathbb{P}(Z_1 \leq \gamma) - \mathbb{P}(Z_1 \leq \gamma)^2\big) \\
&\quad + f_{Z_1}(\gamma)(2\gamma - \|\boldsymbol{\beta}\|)\mathbb{E}\Big(\big(g_1(\|\boldsymbol{Z}\|)Z_1^2 - h_1(\|\boldsymbol{Z}\|)\big)\mathbb{1}\{Z_1 \leq \gamma\}\Big) \\
&\quad + 2w_1^{-1} f_{Z_1}(\gamma)(\gamma/\|\boldsymbol{\beta}\|)\mathbb{E}\big(Z_1 \mathbb{1}\{Z_1 \leq \gamma\}\big)
\end{aligned}
$$

and

$$
\begin{aligned}
\Lambda_{1,2} &= \Lambda_{2,1} \\
&= \Lambda_{1,1} + f_{Z_1}(\gamma)(\gamma - \|\boldsymbol{\beta}\|/2)\mathbb{E}\Big(\big(g_1(\|\boldsymbol{Z}\|)Z_1^2 - h_1(\|\boldsymbol{Z}\|)\big)\mathbb{1}\{Z_1 \leq \gamma\}\Big) \\
&\quad + w_1^{-1} f_{Z_1}(\gamma)(\gamma/\|\boldsymbol{\beta}\|)\mathbb{E}\big(Z_1 \mathbb{1}\{Z_1 \leq \gamma\}\big).
\end{aligned}
$$

We used the dominated convergence theorem twice.

As to Lindeberg's condition, note that $\|\boldsymbol{Y}_{n,i}^{\boldsymbol{\Sigma}}\| \leq n^{-1/2}M$ for some constant $M < \infty$. Thus the proof is similar to the one used for Theorem 3.9. □

3.4. A Central Limit Theorem for Inclusion Probabilities

In this section we consider p-values based on the plug-in statistic for the standard model with two classes. The corresponding conditional inclusion probabilities

$$\mathcal{I}_\alpha(b, \theta \mid \mathcal{D}) = \mathbb{P}(\theta \in \widehat{\mathcal{Y}}_\alpha(\boldsymbol{X}, \mathcal{D}) \mid Y = b, \mathcal{D})$$

are of interest to judge the separability of the two classes. However, these theoretic quantities are typically unknown. Therefore we use cross-validation to estimate them. Namely, we compute the empirical conditional inclusion probabilities

$$\widehat{\mathcal{I}}_\alpha(b, \theta) = N_b^{-1} \# \{i \in \mathcal{G}_b \colon \theta \in \widehat{\mathcal{Y}}_\alpha(\boldsymbol{X}_i, \mathcal{D}_i)\}$$

based on cross-validated p-values.

Dümbgen et al. (2008) showed that $\widehat{\mathcal{I}}_\alpha(b, \theta)$ are consistent estimators of $\mathcal{I}_\alpha(b, \theta \mid \mathcal{D})$, (see also Section 1.6). More precisely,

$$\left.\begin{matrix} \mathcal{I}_\alpha(b, \theta \mid \mathcal{D}) \\ \widehat{\mathcal{I}}_\alpha(b, \theta) \end{matrix}\right\} \to_p \mathbb{P}\big(\pi_\theta^*(\boldsymbol{X}) > \alpha \mid Y = b\big),$$

and the limit equals $1 - \alpha$ in case of $b = \theta$.

We now take a closer look at the inclusion probabilities and describe the asymptotic distribution of

$$\sqrt{n} \begin{pmatrix} \mathcal{I}_\alpha(b, \theta \mid \mathcal{D}) - \mathbb{P}\big(\pi_\theta^*(\boldsymbol{X}) > \alpha \mid Y = b\big) \\ \widehat{\mathcal{I}}_\alpha(b, \theta) - \mathbb{P}\big(\pi_\theta^*(\boldsymbol{X}) > \alpha \mid Y = b\big) \end{pmatrix}$$

assuming only elliptically symmetric instead of Gaussian distributions.

Let $E := \big\{\boldsymbol{x} \in \mathbb{R}^d \colon \pi_\theta^*(\boldsymbol{x}) > \alpha\big\}$ and q denote the $(1 - \alpha)$-quantile of $\mathcal{L}(\boldsymbol{v}^\top \boldsymbol{Z})$ for some unit vector $\boldsymbol{v} \in \mathbb{R}^d$. The spherical symmetry of $\mathcal{L}(\boldsymbol{Z})$ implies that q does not depend on $\boldsymbol{v}$.

Theorem 3.11. *Suppose that $\mathcal{L}(\boldsymbol{Z})$ is elliptically symmetric satisfying* (3.1) *and $\mathbb{E}(\|\boldsymbol{Z}\|^2) < \infty$. Either let $\widehat{\boldsymbol{\Sigma}}$ be the standard estimator and $\mathbb{E}(\|\boldsymbol{Z}\|^4) < \infty$, or let $\widehat{\boldsymbol{\Sigma}}$ be an estimator satisfying the assumptions of Lemma 3.8, e.g. the M-estimators defined in Section 1.1.2. Then for the plug-in rule for the standard model with $L = 2$ classes and $\theta \in \{1, 2\}$,*

$$\begin{aligned} \sqrt{n}\Big(\mathcal{I}_\alpha(\theta, \theta \mid \mathcal{D}) - \mathbb{P}\big(\pi_\theta^*(\boldsymbol{X}) > \alpha \mid Y = \theta\big)\Big) &= -\sqrt{n}\big(\widehat{P}_\theta - P_\theta\big)(E) + \mathrm{o_p}(1) \\ &\to_{\mathcal{L}} \mathcal{N}\big(0, w_\theta^{-1}\alpha(1 - \alpha)\big) \end{aligned}$$

and

$$\sqrt{n}\Big(\widehat{\mathcal{I}}_\alpha(\theta, \theta) - \mathbb{P}\big(\pi_\theta^*(\boldsymbol{X}) > \alpha \mid Y = \theta\big)\Big) = \mathrm{o_p}(1). \tag{3.13}$$

For $\theta \neq \lambda \in \{1,2\}$,

$$\begin{aligned}
&\sqrt{n}\begin{pmatrix} \mathcal{I}_\alpha(\lambda,\theta \mid \mathcal{D}) - \mathbb{P}\big(\pi^*_\theta(\boldsymbol{X}) > \alpha \mid Y = \lambda\big) \\ \widehat{\mathcal{I}}_\alpha(\lambda,\theta) - \mathbb{P}\big(\pi^*_\theta(\boldsymbol{X}) > \alpha \mid Y = \lambda\big) \end{pmatrix} \\
&\quad = \sqrt{n}\begin{pmatrix} c\big(\widehat{P}_\theta - P_\theta\big)(E) \\ c\big(\widehat{P}_\theta - P_\theta\big)(E) + \big(\widehat{P}_\lambda - P_\lambda\big)(E) \end{pmatrix} + \mathrm{o_p}(1) \\
&\quad \to_{\mathcal{L}} \mathcal{N}_2\Bigg(\mathbf{0}, \begin{pmatrix} \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 + \sigma_2^2 \end{pmatrix}\Bigg),
\end{aligned}$$

where $\sigma_1^2 = w_\theta^{-1} c^2 \alpha(1-\alpha)$ and $\sigma_2^2 = w_\lambda^{-1} P_\lambda(E)\big(1 - P_\lambda(E)\big)$ with $c = -f_{Z_1}(q - \|\boldsymbol{\beta}\|)/f_{Z_1}(q)$ and $P_\lambda(E) = F_{Z_1}\big(q - \|\boldsymbol{\beta}\|\big)$.

Corollary 3.12. *Suppose that the assumptions of Theorem 3.11 are satisfied. Then for $b, \theta \in \{1,2\}$*

$$\begin{aligned}
\sqrt{n}\big(\widehat{\mathcal{I}}_\alpha(b,\theta) - \mathcal{I}_\alpha(b,\theta \mid \mathcal{D})\big) &= \sqrt{n}\big(\widehat{P}_b - P_b\big)(E) + \mathrm{o_p}(1) \\
&\to_{\mathcal{L}} \mathcal{N}\big(0, w_b^{-1} P_b(E)\big(1 - P_b(E)\big)\big).
\end{aligned}$$

It is remarkable that the term of order $\mathrm{O_p}(n^{-1/2})$ in $\widehat{\mathcal{I}}_\alpha(\theta,\theta) - \mathbb{P}\big(\pi^*_\theta(\boldsymbol{X}) > \alpha \mid Y = \theta\big)$ vanishes. This means that the cross-validated estimator $\widehat{\mathcal{I}}_\alpha(\theta,\theta)$ converges faster to the inclusion probability of the optimal p-value than the theoretic conditional inclusion probability $\mathcal{I}_\alpha(\theta,\theta \mid \mathcal{D})$ does.

The term of order $\mathrm{O_p}(n^{-1/2})$ in $\widehat{\mathcal{I}}_\alpha(\lambda,\theta) - \mathbb{P}\big(\pi^*_\theta(\boldsymbol{X}) > \alpha \mid Y = \theta\big)$ does not vanish for $\lambda \neq \theta$. But it may be written as a sum of two independent summands. The first summand $\mathcal{I}_\alpha(\lambda,\theta \mid \mathcal{D})$ depends only on the training data from class λ and the second summand depends only on the training data from class θ.

The Corollary implies that the empirical conditional inclusion probabilities $\widehat{\mathcal{I}}_\alpha(b,\theta)$ are root-n-consistent estimators for the conditional inclusion probabilities $\mathcal{I}_\alpha(b,\theta \mid \mathcal{D})$. Moreover, it enables us to construct confidence intervals for $\mathcal{I}_\alpha(b,\theta \mid \mathcal{D})$. An asymptotic $(1-\alpha)$-confidence interval is given by

$$\Big[\widehat{\mathcal{I}}_\alpha(b,\theta) \pm \sqrt{n^{-1} w_b^{-1} P_b(E)\big(1 - P_b(E)\big)}\, z_{1-\alpha/2}\Big],$$

where $z_{1-\alpha/2}$ denotes the $(1-\alpha/2)$-quantile of the standard Gaussian distribution. In practice, the prior probability w_b may be unknown and can be replaced by the deterministic quantity $\widehat{w}_b = N_b/n$. For $b = \theta$, we know that $P_\theta(E) = \alpha$ and can compute the confidence interval

$$\Big[\widehat{\mathcal{I}}_\alpha(\theta,\theta) \pm \sqrt{N_\theta^{-1}\alpha(1-\alpha)}\, z_{1-\alpha/2}\Big]$$

for $\mathcal{I}_\alpha(\theta, \theta \mid \mathcal{D})$. However, for $\lambda \neq \theta$, $P_\lambda(E)$ is typically unknown in practice, which prevents us from constructing confidence intervals for $\mathcal{I}_\alpha(\lambda, \theta \mid \mathcal{D})$.

For the proof of Theorem 3.11 we need the following lemma.

Lemma 3.13. *Suppose that the assumptions of Theorem 3.11 are satisfied. Then for $b, \theta \in \{1, 2\}$*

$$P_b\big(\{\boldsymbol{x} \in \mathbb{R}^d \colon \pi_\theta(\boldsymbol{x}, \mathcal{D}) > \alpha\} \triangle \{\boldsymbol{x} \in \mathbb{R}^d \colon \pi^*_\theta(\boldsymbol{x}) > \alpha\}\big) = \mathrm{O}_\mathrm{p}(n^{-1/2}) \quad (3.14)$$

and

$$\begin{aligned} &P_b\{\boldsymbol{x} \in \mathbb{R}^d \colon \pi_\theta(\boldsymbol{x}, \mathcal{D}) > \alpha\} - P_b\{\boldsymbol{x} \in \mathbb{R}^d \colon \pi^*_\theta(\boldsymbol{x}) > \alpha\} \qquad (3.15)\\ &= \frac{f_{Z_1}(q - \|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_b - \boldsymbol{\mu}_\theta)\|)}{f_{Z_1}(q)} (\widehat{P}_\theta - P_\theta)\{\boldsymbol{x} \in \mathbb{R}^d \colon \pi^*_\theta(\boldsymbol{x}) \leq \alpha\} + \mathrm{o}_\mathrm{p}(n^{-1/2}). \end{aligned}$$

For $b = \theta$ this reduces to

$$\begin{aligned} &P_\theta\{\boldsymbol{x} \in \mathbb{R}^d \colon \pi_\theta(\boldsymbol{x}, \mathcal{D}) > \alpha\} - P_\theta\{\boldsymbol{x} \in \mathbb{R}^d \colon \pi^*_\theta(\boldsymbol{x}) > \alpha\}\\ &\quad = (\widehat{P}_\theta - P_\theta)\{\boldsymbol{x} \in \mathbb{R}^d \colon \pi^*_\theta(\boldsymbol{x}) \leq \alpha\} + \mathrm{o}_\mathrm{p}(n^{-1/2}). \end{aligned}$$

Proof of Lemma 3.13. For $L = 2$ classes and $\theta \neq \lambda \in \{1, 2\}$, we consider

$$T^*_\theta(\boldsymbol{x}) := (\boldsymbol{x} - \boldsymbol{\mu}_{\lambda,\theta})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_\lambda - \boldsymbol{\mu}_\theta),$$

which is a strictly monotonic transformation of $T^*_\theta(\boldsymbol{x})$ defined in Example 1.2 and therefore leads to the same p-values. The empirical version of $T^*_\theta(\boldsymbol{x})$ based on training data $\mathcal{D}$ is given by

$$T_\theta(\boldsymbol{x}, \mathcal{D}) = (\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_{\lambda,\theta})^\top \widehat{\boldsymbol{\Sigma}}^{-1}(\widehat{\boldsymbol{\mu}}_\lambda - \widehat{\boldsymbol{\mu}}_\theta),$$

with estimators $\widehat{\boldsymbol{\mu}}_b(\mathcal{D})$ and $\widehat{\boldsymbol{\Sigma}}^{-1}(\mathcal{D})$. Elementary calculations reveal that $\widehat{\boldsymbol{\mu}}_b(\mathcal{D}_i(\boldsymbol{X})) = \widehat{\boldsymbol{\mu}}_b(\mathcal{D}) + \mathrm{O}_\mathrm{p}(n^{-1})$ and $\widehat{\boldsymbol{\Sigma}}^{-1}(\mathcal{D}_i(\boldsymbol{X})) = \widehat{\boldsymbol{\Sigma}}^{-1}(\mathcal{D}) + \mathrm{O}_\mathrm{p}(n^{-1})$. Thus $T_\theta\big(\boldsymbol{X}_i, \mathcal{D}_i(\boldsymbol{X})\big) = T_\theta(\boldsymbol{X}_i, \mathcal{D}) + \mathrm{O}_\mathrm{p}(n^{-1})$ and

$$\begin{aligned} \pi_\theta(\boldsymbol{X}, \mathcal{D}) &:= \frac{\#\big\{i \in \mathcal{G}_\theta \colon T_\theta\big(\boldsymbol{X}_i, \mathcal{D}_i(\boldsymbol{X})\big) \geq T_\theta(\boldsymbol{X}, \mathcal{D})\big\} + 1}{N_\theta + 1}\\ &= \widehat{P}_\theta\big\{\boldsymbol{z} \in \mathbb{R}^d \colon T_\theta(\boldsymbol{z}, \mathcal{D}) + \mathrm{O}_\mathrm{p}(n^{-1}) \geq T_\theta(\boldsymbol{X}, \mathcal{D})\big\} + \mathrm{O}_\mathrm{p}(n^{-1}). \end{aligned}$$

Consequently,

$$\begin{aligned} &P_b\{\boldsymbol{x} \in \mathbb{R}^d \colon \pi_\theta(\boldsymbol{x}, \mathcal{D}) > \alpha\} - P_b\{\boldsymbol{x} \in \mathbb{R}^d \colon \pi^*_\theta(\boldsymbol{x}) > \alpha\}\\ &\quad = P_b\Big\{\boldsymbol{x} \in \mathbb{R}^d \colon \widehat{P}_\theta\big\{\boldsymbol{z} \in \mathbb{R}^d \colon T_\theta(\boldsymbol{z}, \mathcal{D}) + \mathrm{O}_\mathrm{p}(n^{-1}) \geq T_\theta(\boldsymbol{x}, \mathcal{D})\big\}\\ &\qquad\quad + \mathrm{O}_\mathrm{p}(n^{-1}) > \alpha\Big\} - P_b\Big\{\boldsymbol{x} \in \mathbb{R}^d \colon P_\theta\big\{\boldsymbol{z} \in \mathbb{R}^d \colon T^*_\theta(\boldsymbol{z}) \geq T^*_\theta(\boldsymbol{x})\big\} > \alpha\Big\}\\ &\quad = P_b\big\{\boldsymbol{x} \in \mathbb{R}^d \colon \widehat{P}_\theta(\widehat{A}(\boldsymbol{x})) + \mathrm{O}_\mathrm{p}(n^{-1}) > \alpha\big\} - P_b\big\{\boldsymbol{x} \in \mathbb{R}^d \colon P_\theta(A(\boldsymbol{x})) > \alpha\big\} \end{aligned}$$

where

$$\begin{aligned}\widehat{A}(\boldsymbol{x}) &:= \big\{\boldsymbol{z} \in \mathbb{R}^d \colon T_\theta(\boldsymbol{z}, \mathcal{D}) + \mathrm{O_p}(n^{-1}) \geq T_\theta(\boldsymbol{x}, \mathcal{D})\big\},\\ A(\boldsymbol{x}) &:= \big\{\boldsymbol{z} \in \mathbb{R}^d \colon T^*_\theta(\boldsymbol{z}) \geq T^*_\theta(\boldsymbol{x})\big\}.\end{aligned}$$

Now we split the term in two summands

$$\begin{aligned}&P_b\big\{\boldsymbol{x} \in \mathbb{R}^d \colon \widehat{P}_\theta(\widehat{A}(\boldsymbol{x})) + \mathrm{O_p}(n^{-1}) > \alpha\big\} - P_b\big\{\boldsymbol{x} \in \mathbb{R}^d \colon P_\theta(A(\boldsymbol{x})) > \alpha\big\}\\ &\quad = \Big(P_b\big\{\boldsymbol{x} \in \mathbb{R}^d \colon P_\theta(\widehat{A}(\boldsymbol{x})) > \alpha\big\} - P_b\big\{\boldsymbol{x} \in \mathbb{R}^d \colon P_\theta(A(\boldsymbol{x})) > \alpha\big\}\Big)\\ &\qquad + \Big(P_b\big\{\boldsymbol{x} \in \mathbb{R}^d \colon \widehat{P}_\theta(\widehat{A}(\boldsymbol{x})) + \mathrm{O_p}(n^{-1}) > \alpha\big\}\\ &\qquad\quad - P_b\big\{\boldsymbol{x} \in \mathbb{R}^d \colon P_\theta(\widehat{A}(\boldsymbol{x})) > \alpha\big\}\Big).\end{aligned} \tag{3.16}$$

Regarding the first summand note that

$$\begin{aligned}A(\boldsymbol{x}) &= \big\{\boldsymbol{z} \in \mathbb{R}^d \colon T^*_\theta(\boldsymbol{z}) \geq T^*_\theta(\boldsymbol{x})\big\}\\ &= \big\{\boldsymbol{z} \in \mathbb{R}^d \colon \boldsymbol{z}^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_\lambda - \boldsymbol{\mu}_\theta) \geq \boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_\lambda - \boldsymbol{\mu}_\theta)\big\}\\ &= \big\{\boldsymbol{z} \in \mathbb{R}^d \colon \boldsymbol{\beta}^\top \boldsymbol{\Sigma}^{-1/2}\boldsymbol{z} \geq \boldsymbol{\beta}^\top \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}\big\}\\ &= \big\{\boldsymbol{z} \in \mathbb{R}^d \colon \boldsymbol{u}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{z} - \boldsymbol{\mu}_\theta) \geq \boldsymbol{u}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}_\theta)\big\},\end{aligned}$$

where $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_\lambda - \boldsymbol{\mu}_\theta)$ and $\boldsymbol{u} := \boldsymbol{\beta}/\|\boldsymbol{\beta}\|$. Note that $P_\theta(A(\boldsymbol{x})) > \alpha$ if and only if $\boldsymbol{u}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}_\theta) < q$. Consequently,

$$P_b\big\{\boldsymbol{x} \in \mathbb{R}^d \colon P_\theta(A(\boldsymbol{x})) > \alpha\big\} = P_b\big\{\boldsymbol{x} \in \mathbb{R}^d \colon \boldsymbol{u}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}_\theta) < q\big\}. \tag{3.17}$$

Lemma 3.6, 3.7 and 3.8 yield

$$\begin{aligned}T_\theta(\boldsymbol{x}, \mathcal{D}) &= (\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_{\lambda,\theta})^\top \widehat{\boldsymbol{\Sigma}}^{-1}(\widehat{\boldsymbol{\mu}}_\lambda - \widehat{\boldsymbol{\mu}}_\theta)\\ &= (\boldsymbol{x} - \boldsymbol{\mu}_{\lambda,\theta})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_\lambda - \boldsymbol{\mu}_\theta) - \frac{\boldsymbol{\Delta}^\top_{\boldsymbol{\mu}_{\lambda,\theta}}}{\sqrt{n}} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_\lambda - \boldsymbol{\mu}_\theta)\\ &\quad + (\boldsymbol{x} - \boldsymbol{\mu}_{\lambda,\theta})^\top \frac{\boldsymbol{\Delta}_{\boldsymbol{\Sigma}^{-1}}}{\sqrt{n}}(\boldsymbol{\mu}_\lambda - \boldsymbol{\mu}_\theta) + (\boldsymbol{x} - \boldsymbol{\mu}_{\lambda,\theta})^\top \boldsymbol{\Sigma}^{-1} \frac{\boldsymbol{\Delta}_{\boldsymbol{\mu}_\lambda} - \boldsymbol{\Delta}_{\mu_\theta}}{\sqrt{n}}\\ &\quad + \mathrm{O_p}(n^{-1})\end{aligned}$$

and thus

$$\begin{aligned}
\widehat{A}(\boldsymbol{x}) &= \big\{\boldsymbol{z} \in \mathbb{R}^d \colon T_\theta(\boldsymbol{z}, \mathcal{D}) + \mathrm{O_p}(n^{-1}) \geq T_\theta(\boldsymbol{x}, \mathcal{D})\big\} \\
&= \Big\{\boldsymbol{z} \in \mathbb{R}^d \colon T_\theta^*(\boldsymbol{z}) - T_\theta^*(\boldsymbol{x}) + (\boldsymbol{z}-\boldsymbol{x})^\top \boldsymbol{\Sigma}^{-1/2}\Big(\boldsymbol{\Sigma}^{1/2}\frac{\boldsymbol{\Delta}_{\boldsymbol{\Sigma}^{-1}}}{\sqrt{n}}(\boldsymbol{\mu}_\lambda - \boldsymbol{\mu}_\theta) \\
&\qquad\qquad + \boldsymbol{\Sigma}^{-1/2}\frac{\boldsymbol{\Delta}_{\boldsymbol{\mu}_\lambda} - \boldsymbol{\Delta}_{\boldsymbol{\mu}_\theta}}{\sqrt{n}} + \mathrm{O_p}(n^{-1})\Big) \geq 0\Big\} \\
&= \Big\{\boldsymbol{z} \in \mathbb{R}^d \colon (\boldsymbol{z}-\boldsymbol{x})^\top \boldsymbol{\Sigma}^{-1/2}\Big(\boldsymbol{\beta} + \boldsymbol{\Sigma}^{1/2}\frac{\boldsymbol{\Delta}_{\boldsymbol{\Sigma}^{-1}}}{\sqrt{n}}(\boldsymbol{\mu}_\lambda - \boldsymbol{\mu}_\theta) \\
&\qquad\qquad + \boldsymbol{\Sigma}^{-1/2}\frac{\boldsymbol{\Delta}_{\boldsymbol{\mu}_\lambda} - \boldsymbol{\Delta}_{\boldsymbol{\mu}_\theta}}{\sqrt{n}} + \mathrm{O_p}(n^{-1})\Big) \geq 0\Big\} \\
&= \big\{\boldsymbol{z} \in \mathbb{R}^d \colon \widehat{\boldsymbol{\beta}}^\top \boldsymbol{\Sigma}^{-1/2}\boldsymbol{z} \geq \widehat{\boldsymbol{\beta}}^\top \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}\big\} \\
&= \big\{\boldsymbol{z} \in \mathbb{R}^d \colon \widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{z} - \boldsymbol{\mu}_\theta) \geq \widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}_\theta)\big\},
\end{aligned}$$

where $\widehat{\boldsymbol{\beta}} := \boldsymbol{\beta} + \boldsymbol{\Sigma}^{1/2}\frac{\boldsymbol{\Delta}_{\boldsymbol{\Sigma}^{-1}}}{\sqrt{n}}(\boldsymbol{\mu}_\lambda - \boldsymbol{\mu}_\theta) + \boldsymbol{\Sigma}^{-1/2}\frac{\boldsymbol{\Delta}_{\boldsymbol{\mu}_\lambda} - \boldsymbol{\Delta}_{\boldsymbol{\mu}_\theta}}{\sqrt{n}} + \mathrm{O_p}(n^{-1})$ and $\widehat{\boldsymbol{u}} := \widehat{\boldsymbol{\beta}}/\|\widehat{\boldsymbol{\beta}}\|$.

We deduce from the spherically symmetric distribution of $\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_\theta - \boldsymbol{\mu}_\theta)$ that $\widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_\theta - \boldsymbol{\mu}_\theta)$ conditional on the training data has the same distribution as $\boldsymbol{u}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_\theta - \boldsymbol{\mu}_\theta)$. Therefore $P_\theta\big\{\boldsymbol{z} \in \mathbb{R}^d \colon T_\theta(\boldsymbol{z}, \mathcal{D}) + \mathrm{O_p}(n^{-1}) \geq T_\theta(\boldsymbol{x}, \mathcal{D})\big\} > \alpha$ if and only if $\widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}_\theta) < q$.

This entails that

$$P_b\big\{\boldsymbol{x} \in \mathbb{R}^d \colon P_\theta(\widehat{A}(\boldsymbol{x})) > \alpha\big\} = P_b\big\{\boldsymbol{x} \in \mathbb{R}^d \colon \widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}_\theta) < q\big\}$$

and

$$\begin{aligned}
&P_b\big\{\boldsymbol{x} \in \mathbb{R}^d \colon P_\theta(\widehat{A}(\boldsymbol{x})) > \alpha\big\} - P_b\big\{\boldsymbol{x} \in \mathbb{R}^d \colon P_\theta(A(\boldsymbol{x})) > \alpha\big\} \\
&\quad = P_b\big\{\boldsymbol{x} \in \mathbb{R}^d \colon \widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}_\theta) < q\big\} \\
&\qquad - P_b\big\{\boldsymbol{x} \in \mathbb{R}^d \colon \boldsymbol{u}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}_\theta) < q\big\} \\
&\quad = P_b\big\{\boldsymbol{x} \in \mathbb{R}^d \colon \widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}_b) < \widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_b) + q\big\} \qquad (3.18) \\
&\qquad - P_b\big\{\boldsymbol{x} \in \mathbb{R}^d \colon \boldsymbol{u}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}_b) < \boldsymbol{u}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_b) + q\big\}.
\end{aligned}$$

Next we consider the second summand of (3.16). For $y \in \mathbb{R}$ let

$$\begin{aligned}
G(y) &:= P_\theta\big\{\boldsymbol{z} \in \mathbb{R}^d \colon \boldsymbol{u}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{z} - \boldsymbol{\mu}_\theta) \geq y\big\} \\
&= P_0\{\boldsymbol{z} \in \mathbb{R}^d \colon \boldsymbol{u}^\top \boldsymbol{z} \geq y\} \\
&= 1 - F_{Z_1}(y)
\end{aligned}$$

and therefore $G'(y) = -f_{Z_1}(y) < 0$.

Note that

$$\begin{aligned} P_\theta(A(\boldsymbol{x})) &= P_\theta\big\{\boldsymbol{z} \in \mathbb{R}^d \colon \boldsymbol{u}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{z} - \boldsymbol{\mu}_\theta) \geq \boldsymbol{u}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}_\theta)\big\} \\ &= G(\boldsymbol{u}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}_\theta)) \end{aligned}$$

and the spherically symmetric distribution of $\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_\theta - \boldsymbol{\mu}_\theta)$ implies

$$\begin{aligned} P_\theta(\widehat{A}(\boldsymbol{x})) &= P_\theta\big\{\boldsymbol{z} \in \mathbb{R}^d \colon \widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{z} - \boldsymbol{\mu}_\theta) \geq \widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}_\theta)\big\} \\ &= P_\theta\big\{\boldsymbol{z} \in \mathbb{R}^d \colon \boldsymbol{u}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{z} - \boldsymbol{\mu}_\theta) \geq \widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}_\theta)\big\} \\ &= G(\widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}_\theta)). \end{aligned}$$

The first order Taylor expansion of G at q is given by

$$\begin{aligned} P_\theta(\widehat{A}(\boldsymbol{x})) &= G(q) + G'(q) \cdot \big(\widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}_\theta) - q\big) \\ &\quad + \mathrm{o}_\mathrm{p}\big(\widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}_\theta) - q\big) \\ &= \alpha + G'(q) \cdot \big(\widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}_\theta) - q\big) \\ &\quad + \mathrm{o}_\mathrm{p}\big(\widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}_\theta) - q\big). \end{aligned}$$

For $\boldsymbol{\beta} \in \mathbb{R}^d$ and $\gamma \in \mathbb{R}$ consider the half-space $H(\boldsymbol{\beta}, \gamma) := \{\boldsymbol{z} \in \mathbb{R}^d \colon \boldsymbol{\beta}^\top \boldsymbol{z} + \gamma \leq 0\}$ and the empirical process $\mathbb{B}_{P_\theta,n}(\boldsymbol{\beta}, \gamma) := \sqrt{n}(\widehat{P}_\theta - P_\theta)\big(H(\boldsymbol{\beta}, \gamma)\big)$. Then

$$\begin{aligned} \widehat{P}_\theta(\widehat{A}(\boldsymbol{x})) &= P_\theta(\widehat{A}(\boldsymbol{x})) + \big(\widehat{P}_\theta - P_\theta\big)(\widehat{A}(\boldsymbol{x})) \\ &= P_\theta(\widehat{A}(\boldsymbol{x})) + n^{-1/2}\mathbb{B}_{P_\theta,n}(-\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{u}}, \widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}) \\ &= \alpha + G'(q) \cdot \big(\widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}_\theta) - q\big) \\ &\quad + n^{-1/2}\mathbb{B}_{P_\theta,n}(-\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{u}}, \widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}) \\ &\quad + \mathrm{o}_\mathrm{p}\big(\widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}_\theta) - q\big). \end{aligned}$$

Next we show that for the computation of

$$P_b\big\{\boldsymbol{x} \in \mathbb{R}^d \colon \widehat{P}_\theta(\widehat{A}(\boldsymbol{x})) + \mathrm{O}_\mathrm{p}(n^{-1}) > \alpha\big\} - P_b\big\{\boldsymbol{x} \in \mathbb{R}^d \colon P_\theta(\widehat{A}(\boldsymbol{x})) > \alpha\big\}$$

it suffices to consider all $\boldsymbol{x}$ such that

$$|\widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}_\theta) - q| < n^{-1/2}c^{-1}\|\mathbb{B}_{P_\theta,n}\|_\infty + \mathrm{O}_\mathrm{p}(n^{-1}) = \mathrm{O}_\mathrm{p}(n^{-1/2}) \tag{3.19}$$

for some constant $c > 0$. To this end note that

$$\mathbb{1}\big\{\widehat{P}_\theta(\widehat{A}(\boldsymbol{x})) + \mathrm{O}_\mathrm{p}(n^{-1}) > \alpha\big\} \neq \mathbb{1}\big\{P_\theta(\widehat{A}(\boldsymbol{x})) > \alpha\big\}$$

implies

$$\begin{aligned}\|\mathbb{B}_{P_\theta,n}\|_\infty &> \sqrt{n}\big|P_\theta(\widehat{A}(\boldsymbol{x})) - \alpha\big| + \mathrm{O_p}(n^{-1/2}) \\ &= \sqrt{n}\big|G(\widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}_\theta)) - \alpha\big| + \mathrm{O_p}(n^{-1/2}).\end{aligned}$$

Now suppose that $\big|\widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}_\theta) - q\big| \le \delta$. Because $|G(q) - \alpha\big| = 0$,

$$\big|G(\widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}_\theta)) - \alpha\big| \ge \big|\widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}_\theta) - q\big| \min_{t \in [q \pm \delta]} |G'(t)|,$$

which entails (3.19) with $c := \min_{t\in[q\pm\delta]} |G'(t)|$.

If $\big|\widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}_\theta) - q\big| > \delta$, the monotonicity of G implies that

$$\big|G(\widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}_\theta)) - \alpha\big| \ge \delta c$$

and hence $\|\mathbb{B}_{P_\theta,n}\|_\infty > \sqrt{n}\delta c + \mathrm{O_p}(n^{-1/2})$. But $\|\mathbb{B}_{P_\theta,n}\|_\infty = \mathrm{O_p}(1)$ by Theorem 3.4 and thus $\mathbb{P}(\|\mathbb{B}_{P_\theta,n}\|_\infty > \sqrt{n}\delta c + \mathrm{O_p}(n^{-1/2})) \to 0$.

Suppose that (3.19) holds. Then

$$P_\theta(\widehat{A}(\boldsymbol{x})) = \alpha + G'(q) \cdot \big(\widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}_\theta) - q\big) + \mathrm{o_p}(n^{-1/2}).$$

Moreover, $\widehat{\boldsymbol{u}} = \boldsymbol{u} + \mathrm{O_p}(n^{-1/2})$ and

$$\begin{aligned}\widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x} &= \widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}_\theta) + \widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}_\theta \\ &= q + \boldsymbol{u}^\top \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}_\theta + \mathrm{O_p}(n^{-1/2}).\end{aligned}$$

The spherical symmetry of $\mathcal{L}(\boldsymbol{Z})$ and $\mathbb{E}(\|\boldsymbol{Z}\|^2) < \infty$ imply (3.2). Therefore the assumptions of Lemma 3.1 are satisfied. Applying the lemma, we get

$$\begin{aligned}&P_\theta\big(H(-\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{u}}, \widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}) \triangle H(-\boldsymbol{\Sigma}^{-1/2}\boldsymbol{u}, q + \boldsymbol{u}^\top \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}_\theta)\big) \\ &\quad = \mathrm{O_p}(n^{-1/2})\end{aligned}$$

and by Theorem 3.5,

$$\mathbb{B}_{P_\theta,n}(-\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{u}}, \widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}) = \mathbb{B}_{P_\theta,n}(-\boldsymbol{\Sigma}^{-1/2}\boldsymbol{u}, q + \boldsymbol{u}^\top \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}_\theta) + \mathrm{o_p}(1).$$

Consequently,

$$\begin{aligned}\widehat{P}_\theta(\widehat{A}(\boldsymbol{x})) = \alpha &+ G'(q) \cdot \big(\widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}_\theta) - q\big) \\ &+ n^{-1/2}\mathbb{B}_{P_\theta,n}(-\boldsymbol{\Sigma}^{-1/2}\boldsymbol{u}, q + \boldsymbol{u}^\top \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}_\theta) + \mathrm{o_p}(n^{-1/2}).\end{aligned}$$

Moreover, $G'(q) < 0$ yields

$$\begin{aligned}
&P_b\big\{\boldsymbol{x} \in \mathbb{R}^d \colon \widehat{P}_\theta(\widehat{A}(\boldsymbol{x})) + \mathrm{O}_{\mathrm{p}}(n^{-1}) > \alpha\big\} - P_b\big\{\boldsymbol{x} \in \mathbb{R}^d \colon P_\theta(\widehat{A}(\boldsymbol{x})) > \alpha\big\} \\
&\quad = P_b\big\{\boldsymbol{x} \in \mathbb{R}^d \colon G'(q) \cdot \big(\widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}_\theta) - q\big) \\
&\qquad\qquad + n^{-1/2}\mathbb{B}_{P_\theta,n}(-\boldsymbol{\Sigma}^{-1/2}\boldsymbol{u}, q + \boldsymbol{u}^\top \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}_\theta) + \mathrm{o}_{\mathrm{p}}(n^{-1/2}) > 0\big\} \\
&\qquad - P_b\big\{\boldsymbol{x} \in \mathbb{R}^d \colon P_\theta(\widehat{A}(\boldsymbol{x})) > \alpha\big\} \\
&\quad = P_b\big\{\boldsymbol{x} \in \mathbb{R}^d \colon \widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x} < \widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}_\theta + q \\
&\qquad\qquad - \frac{1}{\sqrt{n}G'(q)}\mathbb{B}_{P_\theta,n}(-\boldsymbol{\Sigma}^{-1/2}\boldsymbol{u}, q + \boldsymbol{u}^\top \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}_\theta) + \mathrm{o}_{\mathrm{p}}(n^{-1/2})\big\} \\
&\qquad - P_b\big\{\boldsymbol{x} \in \mathbb{R}^d \colon P_\theta(\widehat{A}(\boldsymbol{x})) > \alpha\big\} \\
&\quad = P_b\big\{\boldsymbol{x} \in \mathbb{R}^d \colon \widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}_b) < \widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_b) + q \qquad (3.20) \\
&\qquad\qquad - \frac{1}{\sqrt{n}G'(q)}\mathbb{B}_{P_\theta,n}(-\boldsymbol{\Sigma}^{-1/2}\boldsymbol{u}, q + \boldsymbol{u}^\top \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}_\theta) + \mathrm{o}_{\mathrm{p}}(n^{-1/2})\big\} \\
&\qquad - P_b\big\{\boldsymbol{x} \in \mathbb{R}^d \colon P_\theta(\widehat{A}(\boldsymbol{x})) > \alpha\big\}.
\end{aligned}$$

Combining (3.18) and (3.20) we get

$$\begin{aligned}
&P_b\big\{\boldsymbol{x} \in \mathbb{R}^d \colon \widehat{P}_\theta(\widehat{A}(\boldsymbol{x})) + \mathrm{O}_{\mathrm{p}}(n^{-1}) > \alpha\big\} - P_b\big\{\boldsymbol{x} \in \mathbb{R}^d \colon P_\theta(A(\boldsymbol{x})) > \alpha\big\} \\
&\quad = P_b\big\{\boldsymbol{x} \in \mathbb{R}^d \colon \widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}_b) < \widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_b) + q \\
&\qquad\qquad - \frac{1}{\sqrt{n}G'(q)}\mathbb{B}_{P_\theta,n}(-\boldsymbol{\Sigma}^{-1/2}\boldsymbol{u}, q + \boldsymbol{u}^\top \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}_\theta) + \mathrm{o}_{\mathrm{p}}(n^{-1/2})\big\} \\
&\qquad - P_b\big\{\boldsymbol{x} \in \mathbb{R}^d \colon \boldsymbol{u}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}_b) < \boldsymbol{u}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_b) + q\big\} \\
&\quad = P_0\big\{\boldsymbol{x} \in \mathbb{R}^d \colon \widehat{\boldsymbol{u}}^\top \boldsymbol{x} < \widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_b) + q \\
&\qquad\qquad - \frac{1}{\sqrt{n}G'(q)}\mathbb{B}_{P_\theta,n}(-\boldsymbol{\Sigma}^{-1/2}\boldsymbol{u}, q + \boldsymbol{u}^\top \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}_\theta) + \mathrm{o}_{\mathrm{p}}(n^{-1/2})\big\} \\
&\qquad - P_0\big\{\boldsymbol{x} \in \mathbb{R}^d \colon \boldsymbol{u}^\top \boldsymbol{x} < \boldsymbol{u}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_b) + q\big\}
\end{aligned}$$

with $P_0 := \mathcal{L}(\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_b - \boldsymbol{\mu}_b))$. The spherical symmetry of P_0 yields

$$\begin{aligned}
&P_b\big\{\boldsymbol{x} \in \mathbb{R}^d \colon \widehat{P}_\theta(\widehat{A}(\boldsymbol{x})) + \mathrm{O}_{\mathrm{p}}(n^{-1}) > \alpha\big\} - P_b\big\{\boldsymbol{x} \in \mathbb{R}^d \colon P_\theta(A(\boldsymbol{x})) > \alpha\big\} \\
&\quad = P_0\big\{\boldsymbol{x} \in \mathbb{R}^d \colon x_1 < \widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_b) + q \\
&\qquad\qquad - \frac{1}{\sqrt{n}G'(q)}\mathbb{B}_{P_\theta,n}(-\boldsymbol{\Sigma}^{-1/2}\boldsymbol{u}, q + \boldsymbol{u}^\top \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}_\theta) + \mathrm{o}_{\mathrm{p}}(n^{-1/2})\big\} \\
&\qquad - P_0\big\{\boldsymbol{x} \in \mathbb{R}^d \colon x_1 < \boldsymbol{u}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_b) + q\big\} \\
&\quad = P_0\big\{\boldsymbol{x} \in \mathbb{R}^d \colon x_1 - \widehat{s} \le 0\big\} - P_0\big\{\boldsymbol{x} \in \mathbb{R}^d \colon x_1 - s \le 0\big\}
\end{aligned}$$

with $\widehat{s} := \widehat{\boldsymbol{u}}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_b) + q - \frac{1}{\sqrt{n}G'(q)}\mathbb{B}_{P_\theta,n}(-\boldsymbol{\Sigma}^{-1/2}\boldsymbol{u}, q + \boldsymbol{u}^\top \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}_\theta) + \mathrm{o}_{\mathrm{p}}(n^{-1/2})$ and $s := \boldsymbol{u}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_b) + q = q - \|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_b - \boldsymbol{\mu}_\theta)\|$.

Since $\widehat{s} - s = \mathrm{O_p}(n^{-1/2})$ we can apply Lemma 3.1 with $\widehat{\boldsymbol{b}} = \boldsymbol{b} = \boldsymbol{e}_1$, $\widehat{a} = -\widehat{s}$ and $a = -s$. We obtain claim (3.14) and

$$\begin{aligned} &P_b\big\{\boldsymbol{x} \in \mathbb{R}^d \colon \widehat{P}_\theta(\widehat{A}(\boldsymbol{x})) + \mathrm{O_p}(n^{-1}) > \alpha\big\} - P_b\big\{\boldsymbol{x} \in \mathbb{R}^d \colon P_\theta(A(\boldsymbol{x})) > \alpha\big\} \\ &\quad = (\widehat{s} - s) f_{Z_1}(s) + \mathrm{O_p}(n^{-1}). \end{aligned}$$

If $b = \theta$, $(\widehat{\boldsymbol{u}} - \boldsymbol{u})^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_b) = 0$. Suppose that $b \neq \theta$. The first order Taylor expansion of $\|\widehat{\boldsymbol{\beta}}\|$ is then given by

$$\|\widehat{\boldsymbol{\beta}}\| = \|\boldsymbol{\beta}\| + \frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \boldsymbol{\beta}}{\|\boldsymbol{\beta}\|} + o(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|)$$

and therefore

$$\begin{aligned} \widehat{\boldsymbol{u}} - \boldsymbol{u} &= \frac{\|\boldsymbol{\beta}\|\widehat{\boldsymbol{\beta}} - \|\widehat{\boldsymbol{\beta}}\|\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|\|\widehat{\boldsymbol{\beta}}\|} \\ &= \frac{\|\boldsymbol{\beta}\|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \boldsymbol{\beta}}{\|\boldsymbol{\beta}\|}\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|^2 + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \boldsymbol{\beta}} + o(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|) \\ &= \frac{\|\boldsymbol{\beta}\|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \boldsymbol{\beta}}{\|\boldsymbol{\beta}\|}\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|^2} + o(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|) \\ &= \frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\|\boldsymbol{\beta}\|} - \frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \boldsymbol{\beta}}{\|\boldsymbol{\beta}\|^3}\boldsymbol{\beta} + o(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|). \end{aligned}$$

Plugging in $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_\lambda - \boldsymbol{\mu}_\theta)$ and $\widehat{\boldsymbol{\beta}} := \boldsymbol{\beta} + \boldsymbol{\Sigma}^{1/2}\frac{\boldsymbol{\Delta}_{\boldsymbol{\Sigma}^{-1}}}{\sqrt{n}}(\boldsymbol{\mu}_\lambda - \boldsymbol{\mu}_\theta) + \boldsymbol{\Sigma}^{-1/2}\frac{\boldsymbol{\Delta}_{\boldsymbol{\mu}_\lambda} - \boldsymbol{\Delta}_{\boldsymbol{\mu}_\theta}}{\sqrt{n}} + \mathrm{O_p}(n^{-1})$ results in

$$\begin{aligned}\widehat{\boldsymbol{u}} - \boldsymbol{u} = {} & \frac{\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Delta}_{\boldsymbol{\Sigma}^{-1}}(\boldsymbol{\mu}_b - \boldsymbol{\mu}_\theta) + \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\Delta}_{\boldsymbol{\mu}_b} - \boldsymbol{\Delta}_{\boldsymbol{\mu}_\theta})}{\sqrt{n}\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_b - \boldsymbol{\mu}_\theta)\|} \\ & - \frac{(\boldsymbol{\mu}_b - \boldsymbol{\mu}_\theta)^\top \boldsymbol{\Delta}_{\boldsymbol{\Sigma}^{-1}}(\boldsymbol{\mu}_b - \boldsymbol{\mu}_\theta) + (\boldsymbol{\Delta}_{\boldsymbol{\mu}_b} - \boldsymbol{\Delta}_{\boldsymbol{\mu}_\theta})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_b - \boldsymbol{\mu}_\theta)}{\sqrt{n}\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_b - \boldsymbol{\mu}_\theta)\|^3} \\ & \cdot \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_b - \boldsymbol{\mu}_\theta) + \mathrm{o_p}(n^{-1/2}).\end{aligned}$$

Hence

$$\begin{aligned}&(\widehat{\boldsymbol{u}} - \boldsymbol{u})^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_b) \\ &\quad = \frac{(\boldsymbol{\mu}_b - \boldsymbol{\mu}_\theta)^\top \boldsymbol{\Delta}_{\boldsymbol{\Sigma}^{-1}}(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_b) + (\boldsymbol{\Delta}_{\boldsymbol{\mu}_b} - \boldsymbol{\Delta}_{\boldsymbol{\mu}_\theta})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_b)}{\sqrt{n}\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_b - \boldsymbol{\mu}_\theta)\|} \\ &\qquad + \frac{(\boldsymbol{\mu}_b - \boldsymbol{\mu}_\theta)^\top \boldsymbol{\Delta}_{\boldsymbol{\Sigma}^{-1}}(\boldsymbol{\mu}_b - \boldsymbol{\mu}_\theta) + (\boldsymbol{\Delta}_{\boldsymbol{\mu}_b} - \boldsymbol{\Delta}_{\boldsymbol{\mu}_\theta})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_b - \boldsymbol{\mu}_\theta)}{\sqrt{n}\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_b - \boldsymbol{\mu}_\theta)\|} \\ &\qquad + \mathrm{o_p}(n^{-1/2}) \\ &\quad = \mathrm{o_p}(n^{-1/2})\end{aligned}$$

and

$$\begin{aligned}&P_b\{\boldsymbol{x} \in \mathbb{R}^d \colon \pi_\theta(\boldsymbol{x}, \mathcal{D}) > \alpha\} - P_b\{\boldsymbol{x} \in \mathbb{R}^d \colon \pi^*_\theta(\boldsymbol{x}) > \alpha\} \\ &\quad = (\widehat{s} - s) f_{Z_1}(s) + \mathrm{O_p}(n^{-1}) \\ &\quad = -\frac{1}{\sqrt{n} G'(q)} \mathbb{B}_{P_\theta, n}(-\boldsymbol{\Sigma}^{-1/2}\boldsymbol{u}, q + \boldsymbol{u}^\top \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}_\theta) f_{Z_1}(s) + \mathrm{o_p}(n^{-1/2}).\end{aligned}$$

Finally, with

$$\begin{aligned}H(-\boldsymbol{\Sigma}^{-1/2}\boldsymbol{u}, q + \boldsymbol{u}^\top \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}_\theta) &= \{\boldsymbol{x} \in \mathbb{R}^d \colon \boldsymbol{u}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}_\theta) \geq q\} \\ &= \{\boldsymbol{x} \in \mathbb{R}^d \colon \pi^*_\theta(\boldsymbol{x}) \leq \alpha\}\end{aligned}$$

we get

$$\begin{aligned}&P_b\{\boldsymbol{x} \in \mathbb{R}^d \colon \pi_\theta(\boldsymbol{x}, \mathcal{D}) > \alpha\} - P_b\{\boldsymbol{x} \in \mathbb{R}^d \colon \pi^*_\theta(\boldsymbol{x}) > \alpha\} \\ &\quad = \frac{f_{Z_1}(q - \|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_b - \boldsymbol{\mu}_\theta)\|)}{f_{Z_1}(q)} (\widehat{P}_\theta - P_\theta)\big(\{\boldsymbol{x} \in \mathbb{R}^d \colon \pi^*_\theta(\boldsymbol{x}) \leq \alpha\}\big) \\ &\qquad + \mathrm{o_p}(n^{-1/2}).\end{aligned}$$

□

Proof of Theorem 3.11. Define $\widehat{E} := \{\boldsymbol{x} \in \mathbb{R}^d \colon \pi_\theta(\boldsymbol{x}, \mathcal{D}) > \alpha\}$ and note that $\widehat{\mathcal{I}}_\alpha(b,\theta) = \widehat{P}_b(\widehat{E}) + \mathrm{o}_\mathrm{p}(n^{-1/2})$. Then by Theorem 3.5 and equation (3.14) from Lemma 3.13,

$$\begin{aligned}
&\widehat{\mathcal{I}}_\alpha(b,\theta) - \mathbb{P}\big(\pi^*_\theta(\boldsymbol{X}) > \alpha \mid Y = b\big) \\
&\quad = \widehat{P}_b(\widehat{E}) - P_b(E) + \mathrm{o}_\mathrm{p}(n^{-1/2}) \\
&\quad = \big(P_b(\widehat{E}) - P_b(E)\big) + \big(\widehat{P}_b - P_b\big)(E) \\
&\qquad + \Big(\big(\widehat{P}_b - P_b\big)(\widehat{E}) - \big(\widehat{P}_b - P_b\big)(E)\Big) + \mathrm{o}_\mathrm{p}(n^{-1/2}) \\
&\quad = \big(P_b(\widehat{E}) - P_b(E)\big) + \big(\widehat{P}_b - P_b\big)(E) + \mathrm{o}_\mathrm{p}(n^{-1/2}).
\end{aligned}$$

Let

$$\begin{aligned}
\boldsymbol{\eta} &:= \sqrt{n} \begin{pmatrix} \mathcal{I}_\alpha(b,\theta \mid \mathcal{D}) - \mathbb{P}\big(\pi^*_\theta(\boldsymbol{X}) > \alpha \mid Y = b\big) \\ \widehat{\mathcal{I}}_\alpha(b,\theta) - \mathbb{P}\big(\pi^*_\theta(\boldsymbol{X}) > \alpha \mid Y = b\big) \end{pmatrix} \\
&= \sqrt{n} \begin{pmatrix} P_b(\widehat{E}) - P_b(E) \\ \big(P_b(\widehat{E}) - P_b(E)\big) + \big(\widehat{P}_b - P_b\big)(E) \end{pmatrix} + \mathrm{o}_\mathrm{p}(1).
\end{aligned}$$

We employ Lemma 3.13 to decompose $P_b(\widehat{E}) - P_b(E)$ in independent summands

$$\begin{aligned}
P_b(\widehat{E}) - P_b(E) &= -c\big(\widehat{P}_\theta - P_\theta\big)\{\boldsymbol{x} \in \mathbb{R}^d \colon \pi^*_\theta(\boldsymbol{x}) \le \alpha\} + \mathrm{o}_\mathrm{p}(n^{-1/2}) \\
&= c\big(\widehat{P}_\theta - P_\theta\big)(E) + \mathrm{o}_\mathrm{p}(n^{-1/2}),
\end{aligned}$$

where $c = -f_{Z_1}(q - \|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_b - \boldsymbol{\mu}_\theta)\|)/f_{Z_1}(q)$. Note that $c = -1$ if $b = \theta$. In this case

$$\begin{aligned}
&\widehat{P}_\theta\big\{\boldsymbol{x} \in \mathbb{R}^d \colon \pi_\theta(\boldsymbol{x}, \mathcal{D}) > \alpha\big\} - P_\theta\big\{\boldsymbol{x} \in \mathbb{R}^d \colon \pi^*_\theta(\boldsymbol{x}) > \alpha\big\} \\
&\quad = -\big(\widehat{P}_\theta - P_\theta\big)(E) + \big(\widehat{P}_\theta - P_\theta\big)(E) + \mathrm{o}_\mathrm{p}(n^{-1/2}) \\
&\quad = \mathrm{o}_\mathrm{p}(n^{-1/2}),
\end{aligned}$$

which is claim (3.13).

Now we define for $\nu \in \{b, \theta\}$

$$Y^\nu_{n,i} := \frac{1}{\sqrt{n}\widehat{w}_\nu} \mathbb{1}\{i \in \mathcal{G}_\nu\}\big(\mathbb{1}\{\boldsymbol{X}_i \in E\} - P_\nu(E)\big)$$

such that

$$\begin{aligned}
\sum_{i=1}^n Y^\nu_{n,i} &= \frac{\sqrt{n}}{N_\nu} \sum_{i \in \mathcal{G}_\nu} \big(\mathbb{1}\{\boldsymbol{X}_i \in E\} - P_\nu(E)\big) \\
&= \sqrt{n}\big(\widehat{P}_\nu - P_\nu\big)(E).
\end{aligned}$$

Thus we may write

$$\boldsymbol{\eta} = \sum_{i=1}^{n} \boldsymbol{Y}_{n,i}^{\boldsymbol{\eta}} + \mathrm{o_p}(1)$$

with

$$\boldsymbol{Y}_{n,i}^{\boldsymbol{\eta}} := \begin{pmatrix} cY_{n,i}^{\theta} \\ cY_{n,i}^{\theta} + Y_{n,i}^{b} \end{pmatrix}.$$

Before we can apply the central limit theorem to $\sum_{i=1}^{n} \boldsymbol{Y}_{n,i}^{\boldsymbol{\eta}}$ we need to compute the covariance matrix $\operatorname{Var}(\boldsymbol{Y}_{n,i}^{\boldsymbol{\eta}})$, the sum $\boldsymbol{\Lambda}^n := \sum_{i=1}^{n} \operatorname{Var}(\boldsymbol{Y}_{n,i}^{\boldsymbol{\eta}})$ and its limit $\boldsymbol{\Lambda} = \lim_{n\to\infty} \boldsymbol{\Lambda}^n$. To this end note that

$$\operatorname{Var}(Y_{n,i}^{\nu}) = \frac{\mathbb{1}\{i \in \mathcal{G}_\nu\}}{n\widehat{w}_\nu^2}\big(P_\nu(E) - P_\nu(E)^2\big)$$

and

$$\begin{aligned}
\Lambda_{1,1}^n &= c^2 \sum_{i=1}^{n} \operatorname{Var}(Y_{n,i}^{\theta}) \\
&\to \frac{c^2}{w_\theta}\big(P_\theta(E) - P_\theta(E)^2\big) \\
&= w_\theta^{-1} c^2 \alpha(1-\alpha).
\end{aligned}$$

Suppose that $b \neq \theta$. Then $\operatorname{Cov}(Y_{n,i}^{\theta}, Y_{n,i}^{b}) = 0$ and thus $\Lambda_{1,1} = \Lambda_{1,2} = \Lambda_{2,1}$. Moreover,

$$\begin{aligned}
\operatorname{Var}(cY_{n,i}^{\theta} + Y_{n,i}^{\lambda}) &= c^2 \operatorname{Var}(Y_{n,i}^{\theta}) + \operatorname{Var}(Y_{n,i}^{\lambda}) \\
&= \frac{c^2 \mathbb{1}\{i \in \mathcal{G}_\theta\}}{n\widehat{w}_\theta^2}\big(P_\theta(E) - P_\theta(E)^2\big) \\
&\quad + \frac{\mathbb{1}\{i \in \mathcal{G}_\lambda\}}{n\widehat{w}_\lambda^2}\big(P_\lambda(E) - P_\lambda(E)^2\big)
\end{aligned}$$

and

$$\begin{aligned}
\Lambda_{2,2}^n &= c^2 \widehat{w}_\theta^{-1}\big(P_\theta(E) - P_\theta(E)^2\big) + \widehat{w}_\lambda^{-1}\big(P_\lambda(E) - P_\lambda(E)^2\big) \\
&\to w_\theta^{-1} c^2 \alpha(1-\alpha) + w_\lambda^{-1} P_\lambda(E)\big(1 - P_\lambda(E)\big)
\end{aligned}$$

with

$$\begin{aligned}
P_\lambda(E) &= P_0\big\{\boldsymbol{x} \in \mathbb{R}^d \colon \boldsymbol{u}^\top \boldsymbol{x} < q + \boldsymbol{u}^\top \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_\lambda)\big\} \\
&= P_0\big\{\boldsymbol{x} \in \mathbb{R}^d \colon \boldsymbol{u}^\top \boldsymbol{x} < q - \|\boldsymbol{\beta}\|\big\} \\
&= F_{Z_1}\big(q - \|\boldsymbol{\beta}\|\big),
\end{aligned}$$

which follows from equation (3.17).

As to Lindeberg's condition, note that $|Y^{\nu}_{n,i}| \leq 2n^{-1/2}\widehat{w}_{\nu}^{-1}$ and hence

$$\|\boldsymbol{Y}^{\boldsymbol{\eta}}_{n,i}\| \leq 2|c||Y^{\theta}_{n,i}| + |Y^{b}_{n,i}| \leq \frac{4|c|}{\sqrt{n}\widehat{w}_{\theta}} + \frac{2}{\sqrt{n}\widehat{w}_{b}} \leq \frac{M_n}{\sqrt{n}},$$

where M_n is deterministic and bounded. Therefore Lindeberg's condition

$$\sum_{i=1}^{n} \mathbb{E}\big(\|\boldsymbol{Y}^{\boldsymbol{\eta}}_{n,i}\|^2 \min(1, \|\boldsymbol{Y}^{\boldsymbol{\eta}}_{n,i}\|)\big) \leq n^{-1/2} M_n^3 \to 0$$

is satisfied. Finally note that $\mathbb{E}(\boldsymbol{Y}^{\boldsymbol{\eta}}_{n,i}) = \mathbf{0}$ for all $i \leq n$. Now the assertions follow from the multivariate central limit theorem (Theorem A.1). □

4. Randomized and De-Randomized P-Values

In this chapter we discuss the concept of randomized p-values which is not directly related to the main topic of this thesis. The corresponding concept of randomized tests is familiar in mathematical statistics and is used to obtain tests with exact prescribed significance levels even in settings with test statistics having discrete distributions. Similarly, randomized p-values are particularly useful for test statistics with discrete distributions.

In applications non-randomized tests and p-values are needed. Therefore we review and modify a method of Meinshausen et al. (2009) to de-randomize p-values. To randomize and de-randomize one single p-value brings no benefit, since the resulting p-value is greater than the initial one. However, if we want to combine p-values obtained from different independent test statistics, randomization may be useful. Randomize the p-values, combine them and de-randomize the combination again can lead to a result which is considerably smaller than the combination of the non-randomized p-values.

4.1. De-Randomization

Example 4.1. We consider a random variable $X \sim \text{Poiss}(\lambda)$ with unknown parameter λ and want to test $H_0 \colon \lambda = \lambda_0$ versus $H_A \colon \lambda > \lambda_0$. The usual p-value is given by $1 - G_{\lambda_0}(T-1)$ with the test statistic $T = X$ and G_{λ_0} denoting the c.d.f. of $\text{Poiss}(\lambda_0)$. This p-value is conservative due to the discrete distribution of T.

In mathematical statistics we look at the randomized level-α-test rejecting H_0 with probability

$$\varphi(x) = \begin{cases} 1 & \text{if } \ 1 - G_{\lambda_0}(T-1) \leq \alpha \\ \gamma & \text{if } \ 1 - G_{\lambda_0}(T) \leq \alpha < 1 - G_{\lambda_0}(T-1) \\ 0 & \text{if } \ \alpha < 1 - G_{\lambda_0}(T), \end{cases}$$

where

$$\gamma := \frac{\mathbb{P}\big(1 - G_{\lambda_0}(T-1)\big) - \alpha}{\mathbb{P}\big(1 - G_{\lambda_0}(T-1)\big) - \mathbb{P}\big(1 - G_{\lambda_0}(T)\big)}.$$

The corresponding randomized p-value is given by

$$\pi \sim \mathrm{Unif}[1 - G_{\lambda_0}(T), 1 - G_{\lambda_0}(T-1)].$$

In applications non-randomized tests and p-values are needed. Therefore we will de-randomize this p-value.

The abstract setting. Consider a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, and let $T : (\Omega, \mathcal{A}) \to (\mathcal{T}, \mathcal{B})$ and $\pi : (\Omega, \mathcal{A}) \to [0,1]$ be measurable mappings satisfying the following assumptions under a certain null hypothesis:

(A.1) $\mathcal{L}(\pi) = \mathrm{Unif}[0,1]$.
(A.2) For each $t \in \mathcal{T}$ there exists a given distribution function F_t on $[0,1]$ such that

$$\mathbb{P}(\pi \le u \,|\, T) \;=\; F_T(u) \quad \text{almost surely, for each } u \in [0,1].$$

Note that π is a randomized p-value, and the pair (T, π) may be represented as

$$(T, \pi) \;=\; \big(T, Q_T(U)\big),$$

where T and U are independent, $U \sim \mathrm{Unif}[0,1]$, and Q_t is the quantile function of F_t, i.e.

$$Q_t(v) \;:=\; \min\big\{u \in [0,1] : F_t(u) \ge v\big\}.$$

In our specific applications, T is a single or vector-valued test statistic, and it is often desirable to come up with a p-value $\tilde{\pi}$ depending on T only. A naive solution would be

$$\tilde{\pi} \;:=\; Q_T(1),$$

because $Q_T(1) \ge \pi$ almost surely. But this may be much too conservative. Here is a first general proposal how to construct $\tilde{\pi}$:

De-randomization in the spirit of Meinshausen et al. (2009). Let $\Gamma \subset (0,1]$ be a nonvoid set, and let $h : \Gamma \to (0, \infty)$. Defining

$$J(u) \;:=\; \sup_{\gamma \in \Gamma} \frac{\mathbb{1}\{u \le \gamma\}}{h(\gamma)} \quad \text{for } u \ge 0,$$

we assume that

$$J := \int_0^1 J(u)\, du \;<\; \infty.$$

We define $Q_t(v) := \infty$ for $v > 1$ and note that $J(u) = 0$ for $u > 1$. Then for any $\alpha \in (0,1)$,

$$\begin{aligned}
\alpha &= \mathbb{E}\Big(\frac{J(\pi/\alpha)}{J}\Big) \\
&= \mathbb{E}\Big(\mathbb{E}\Big(\frac{J(\pi/\alpha)}{J}\,\Big|\,T\Big)\Big) \\
&\geq \mathbb{E}\Big(\sup_{\gamma\in\Gamma}\mathbb{E}\Big(\frac{\mathbb{1}\{\pi \leq \alpha\gamma\}}{Jh(\gamma)}\,\Big|\,T\Big)\Big) \\
&= \mathbb{E}\Big(\sup_{\gamma\in\Gamma}\frac{F_T(\alpha\gamma)}{Jh(\gamma)}\Big) \\
&\geq \mathbb{E}\Big(\sup_{\gamma\in\Gamma}\mathbb{1}\big\{F_T(\alpha\gamma) \geq Jh(\gamma)\big\}\Big) \\
&= \mathbb{E}\Big(\sup_{\gamma\in\Gamma}\mathbb{1}\big\{Q_T(Jh(\gamma)) \leq \alpha\gamma\big\}\Big) \\
&= \mathbb{P}\big(Q_T(Jh(\gamma)) \leq \alpha\gamma \text{ for some } \gamma \in \Gamma\big).
\end{aligned}$$

Thus we may reject the null hypothesis at level α if

$$\frac{Q_T(Jh(\gamma))}{\gamma} \;\leq\; \alpha \text{ for some } \gamma \in \Gamma.$$

A corresponding non-randomized p-value is given by

$$\tilde{\pi} \;:=\; \min_{\gamma\in\Gamma}\frac{Q_T(Jh(\gamma))}{\gamma},$$

provided the latter minimum exists almost surely.

Example 4.2. Let $\Gamma = \{\gamma_o\}$ for some fixed $\gamma_o \in (0,1)$ and $h(\gamma_o) = \gamma_o$. Then $J(u) = 1\{u \leq \gamma_o\}/\gamma_o$, $J = 1$, and the nonrandomized p-value is given by

$$\tilde{\pi} \;:=\; \frac{Q_T(\gamma_o)}{\gamma_o}.$$

Example 4.3. Let $\Gamma = [\gamma_o, 1]$ for some fixed $\gamma_o \in (0,1)$ and $h(\gamma) = \gamma$. Then

$$J(u) = \sup_{\gamma_o\leq\gamma\leq 1}\frac{\mathbb{1}\{u \leq \gamma\}}{\gamma} \;=\; \begin{cases} 1/\gamma_o & \text{if } \; 0 \leq u \leq \gamma_o, \\ 1/u & \text{if } \; \gamma_o \leq u \leq 1. \end{cases}$$

Thus

$$J \;=\; \int_0^{\gamma_o}\frac{1}{\gamma_o}\,du + \int_{\gamma_o}^1\frac{1}{u}\,du \;=\; \log(e/\gamma_o).$$

Consequently, a nonrandomized p-value is given by

$$\begin{aligned}\tilde{\pi} &:= \min_{\gamma_o \le \gamma \le 1} \frac{Q_T(\log(e/\gamma_o)\gamma)}{\gamma} \\ &= \min_{\gamma_o \le \gamma \le 1/\log(e/\gamma_o)} \frac{Q_T(\log(e/\gamma_o)\gamma)}{\gamma} \\ &= \min_{\gamma_o \log(e/\gamma_o) \le u \le 1} \frac{\log(e/\gamma_o) Q_T(u)}{u}.\end{aligned}$$

Example 4.4. Let $\Gamma = (0,1]$ and $h(\gamma) = \gamma^\delta$ for some $\delta \in (0,1)$. Then

$$J(u) = \sup_{0<\gamma\le 1} \frac{\mathbb{1}\{u \le \gamma\}}{\gamma^\delta} = \mathbb{1}\{u \le 1\} u^{-\delta} \quad \text{and} \quad J = (1-\delta)^{-1}.$$

Consequently, a nonrandomized test rejects the null hypothesis if

$$\frac{Q_T(\gamma^\delta/(1-\delta))}{\gamma} \le \alpha \quad \text{for some } \gamma \in (0,1],$$

which is equivalent to

$$\frac{Q_T(u)}{((1-\delta)u)^{1/\delta}} \le \alpha \quad \text{for some } u \in (0,1].$$

The corresponding p-value,

$$\tilde{\pi} = \min_{0<u\le 1} \frac{Q_T(u)}{((1-\delta)u)^{1/\delta}},$$

is well-defined if, for instance,

$$\limsup_{u \downarrow 0} \frac{F_t(u)}{u} < \infty \quad \text{for all } t \in \mathcal{T}.$$

For then, $Q_t(u) \ge c(t)u$ for all $u \in (0,1]$ and some $c(t) > 0$, so that $Q_t(u)/u^{1/\delta} \to \infty$ as $u \downarrow 0$. Moreover, Q_t is left-continuous and non-decreasing, and this entails that $Q_t(u)/u^{1/\delta}$ attains a minimum on $(0,1]$.

4.2. Combining Independent P-Values

Suppose that for a given null hypothesis, stochastically independent and possibly randomized p-values $\pi_1, \pi_2, \ldots, \pi_m$ are available. There are infinitely

many possibilities to combine these p-values into one p-value. One specific way is to use

$$\pi \;:=\; \Phi\Big(m^{-1/2}\sum_{i=1}^{m}\Phi^{-1}(\pi_i)\Big) \tag{4.1}$$

with the standard Gaussian distribution function Φ. More generally, we may define

$$\pi \;:=\; \Phi\Big(\sum_{i=1}^{m} w_i\Phi^{-1}(\pi_i)\Big) \tag{4.2}$$

with certain weights $w_1, w_2, \ldots, w_m > 0$ such that $\sum_{i=1}^m w_i^2 = 1$.

This method is motivated by the following model: Suppose we observe independent random variables $X_1, X_2, \ldots, X_m$ with $X_i \sim \mathcal{N}(\theta, \sigma_i^2)$, where the standard deviations $\sigma_i > 0$ are known while the mean θ is unknown. For the null hypothesis "$\theta \geq 0$", possible p-values are given by

$$\pi_i \;:=\; \Phi\Big(\frac{X_i}{\sigma_i}\Big).$$

Then proposal (4.1) leads to

$$\Phi\Big(m^{-1/2}\sum_{i=1}^{m}\frac{X_i}{\sigma_i}\Big) \;\sim\; \mathrm{Unif}[0,1] \quad \text{if } \theta = 0.$$

In this simple model, the Neyman-Pearson Lemma (Theorem A.2) shows that an optimal p-value is given by

$$\Phi\bigg(\Big(\sum_{i=1}^{m}\sigma_i^{-2}\Big)^{-1/2}\sum_{i=1}^{m}\frac{X_i}{\sigma_i^2}\bigg) \;=\; \Phi\bigg(\Big(\sum_{i=1}^{m}\sigma_i^{-2}\Big)^{-1/2}\sum_{i=1}^{m}\frac{\Phi^{-1}(\pi_i)}{\sigma_i}\bigg),$$

and this corresponds to proposal (4.2) with

$$w_i \;:=\; \Big(\sum_{j=1}^{m}\sigma_j^{-2}\Big)^{-1/2}\sigma_i^{-1}.$$

Here the optimal weights are proportional to the reciprocal standard deviations. But there is another representation which can be imitated in different settings: Note that with $Z \sim \mathcal{N}(0,1)$,

$$\frac{d}{d\theta}\Big|_{\theta=0}\mathbb{E}_\theta\pi_i \;=\; \frac{d}{d\theta}\Big|_{\theta=0}\mathbb{E}\Phi\Big(\frac{\theta+\sigma_i Z}{\sigma_i}\Big) \;=\; \frac{1}{\sigma_i}\mathbb{E}\Phi'(Z) \;=\; \frac{\text{const.}}{\sigma_i}.$$

Thus in more general settings where our null hypothesis is of the form "$\theta \geq 0$" for some real parameter θ, we propose to choose w_i proportional to

$$\frac{d}{d\theta}\Big|_{\theta=0}\mathbb{E}_\theta\pi_i.$$

4.3. Application to Multiple Contingency Tables

Suppose that we have multiple independent contingency tables, e.g. data from a multicenter clinical trial or a meta-analysis. Assume that the tables have small cell counts. In this case the test statistics can take only few values and randomization may be useful.

4.3.1. Two-by-Two Tables

Let $S^{(1)}, S^{(2)}, \ldots, S^{(m)}$ be independent two-by-two tables:

X_i	$z_i - X_i$	z_i
$s_i - X_i$	$n_i - s_i - z_i + X_i$	$n_i - z_i$
s_i	$n_i - s_i$	n_i

We consider the row and column sums as fixed and assume that the tables have a common but unknown odds ratio ρ.

Now we want to test the null hypothesis $\rho = 1$ versus the alternative $\rho < 1$. Under the null hypothesis X_i has a hypergeometric distribution with parameters n_i, z_i and s_i. A p-value for table $S^{(i)}$ is given by $\pi_i^{\text{cons}} := H_{n_i,z_i,s_i}(X_i)$ where $H_{n_i,z_i,s_i}(k) := \sum_{l=0}^{k} \left(\binom{z_i}{l}\binom{n-z_i}{s_i-l}\Big/\binom{n_i}{s_i}\right)$ denotes the cdf of the hypergeometric distribution with parameters n_i, z_i and s_i. Due to its discrete distribution, this p-value may be rather conservative. Therefore we consider the randomized p-value

$$\pi_i \sim \text{Unif}[H_{n_i,z_i,s_i}(X_i - 1), H_{n_i,z_i,s_i}(X_i)].$$

Combining the p-values of all tables, we get the randomized p-value

$$\pi \;:=\; \Phi\Big(\sum_{i=1}^{m} w_i \Phi^{-1}(\pi_i)\Big),$$

which we de-randomize as described in section 4.1. Note that the non-randomized p-values π_i^{cons} can be combined the same way, which results in the conservative p-value $\pi^{\text{cons}} := \Phi\big(\sum_{i=1}^{m} w_i \Phi^{-1}(\pi_i^{\text{cons}})\big)$.

In order to compute Q_T with $T = (X_i)_{i=1}^m$, we first approximate F_T numerically. To this end we discretize the distribution of the $Z_i := w_i \Phi^{-1}(\pi_i)$. For a fixed $\delta > 0$ we choose $C_{i,1}, C_{i,2} \in \mathbb{Z}\delta$ such that

$$\mathbb{P}(Z_i \le C_{i,1} \mid T) \ll 1,$$
$$\mathbb{P}(Z_i > C_{i,2} \mid T) \ll 1.$$

Then we define

$$\tilde{Z}_i := \begin{cases} C_{i,1} & \text{if } Z_i \le C_{i,1}, \\ \lceil Z_i/\delta \rceil \delta & \text{if } C_{i,1} < Z_i < C_{i,2}, \\ C_{i,2} & \text{if } Z_i \ge C_{i_2}, \end{cases}$$

and compute $p_i = \big(p_i(j)\big)_{j=1}^{M(i)}$, with $p_i(j) := \mathbb{P}(\tilde{Z}_i = C_{i,1} + (j-1)\delta \mid T)$. The approximate distribution of π conditional on T is then given by the convolution of $p_1, p_2, \ldots, p_m$ and its domain is given by

$$\Phi\Big(\sum_{i=1}^{m} C_{i,1} + k\delta\Big) \qquad k = 0, 1, \ldots, \sum_{i=1}^{m} \big(M(i) - 1\big),$$

where $M(i) := (C_{i,2} - C_{i,1})/\delta + 1$.

Simulation of X_i. To compute the power of the resulting test we need to simulate X_i under the alternative hypothesis. In general X_i has a non-central hypergeometric distribution, i.e.

$$\mathbb{P}(X_i = k \mid n_i, z_i, s_i) = f_\rho(k \mid n_i, z_i, s_i),$$

with

$$f_\rho(k \mid n, z, s) := C_\rho(n, z, s)^{-1} \rho^k \binom{z}{k} \binom{n-z}{s-k},$$

$$C_\rho(n, z, s) := \sum_{l=\max(z+s-n,0)}^{\min(z,s)} \rho^l \binom{z}{l} \binom{n-z}{s-l},$$

see e.g. Agresti (2007). For the computation we use the representation

$$f_\rho(k \mid n, z, s) = \tilde{C}_\rho(n, z, s)^{-1} \frac{\rho^k}{k!(z-k)!(s-k)!(n-z-s+k)!},$$

$$\tilde{C}_\rho(n, z, s) := \sum_{l=\max(z+s-n,0)}^{\min(z,s)} \frac{\rho^l}{l!(z-l)!(s-l)!(n-z-s+l)!}$$

and to avoid numerical problems, we compute it in three steps:

$$\begin{aligned} f &\leftarrow \log(\rho)k - \log(k!) - \log\big((z-k)!\big) - \log\big((s-k)!\big) \\ &\qquad - \log\big((n-s-z+k)!\big) \\ f &\leftarrow \exp\big((f - \max(f))\big) \\ f &\leftarrow f/\mathrm{sum}(f). \end{aligned}$$

	$\rho = 1$	$\rho = 0.8$	$\rho = 0.5$	$\rho = 1/3$
exact	0.0355	0.1443	0.7287	0.9809
conservative	0.0013	0.0147	0.2949	0.8189
$\beta = 0.1$	0.0064	0.0518	0.5204	0.9298
$\beta = 0.3$	0.0135	0.0907	0.6328	0.9615
$\beta = 0.5$	0.0135	0.0907	0.6328	0.9615
$\beta = 0.7$	0.0135	0.0907	0.6328	0.9615
$\beta = 0.9$	0.0135	0.0907	0.6328	0.9615
$\gamma_0 = 0.5$	0.0064	0.0518	0.5204	0.9298
$\gamma_0 = 0.1$	0.0031	0.0280	0.4064	0.8830

Table 4.1.: Power for Example 4.5 at significance level $\alpha = 0.05$.

Exact Monte-Carlo p-values. In the case of two-by-two tables we can compute exact Monte-Carlo p-values. We use them as a benchmark for the de-randomized p-values. In practice the de-randomized p-values are only useful for K-by-L tables with $\max(K, L) > 2$.

For the computation of the Monte-Carlo p-values note that

$$\log \mathbb{P}_\theta(X_i = k \mid n_i, z_i, s_i) = k\theta + \log\left(C_\rho(n, z, s)^{-1}\right) + \log\left(\binom{z}{k}\binom{n-z}{s-k}\right)$$

with $\theta := \log \rho$. The log-likelihood function of the whole model is given by

$$L(\theta) = \sum_{i=1}^{m} \left(\theta X_i + \log\left(\binom{z}{X_i}\binom{n-z}{s-X_i}\right)\right) + m \log\left(C_\rho(n, z, s)^{-1}\right)$$

and a potential test statistic for "$\theta = \theta_0$" vs. "$\theta < \theta_0$" would be

$$\frac{\partial}{\partial\theta}\Big|_{\theta=\theta_0} L(\theta) = \sum_{i=1}^{m} X_i + m \frac{\partial}{\partial\theta}\Big|_{\theta=\theta_0} \log\left(C_\rho(n, z, s)^{-1}\right).$$

Since the last summand does not depend on the data, we choose the test statistic

$$\sum_{i=1}^{m} X_i.$$

Example 4.5. We simulated 10'000 times $m = 10$ tables with $n = 20$, $z = 10$, $s = 8$ and different values for the odds ratio ρ. To combine the p-values we used equal weights $w_i = m^{-1/2}$. Table 4.1 shows the power at significance level $\alpha = 0.05$ for the exact Monte-Carlo p-value, the conservative p-value π^{cons}, $Q_T(\beta)/\beta$ for $\beta = 0.1, 0.3, \ldots, 0.9$ and the adaptive version

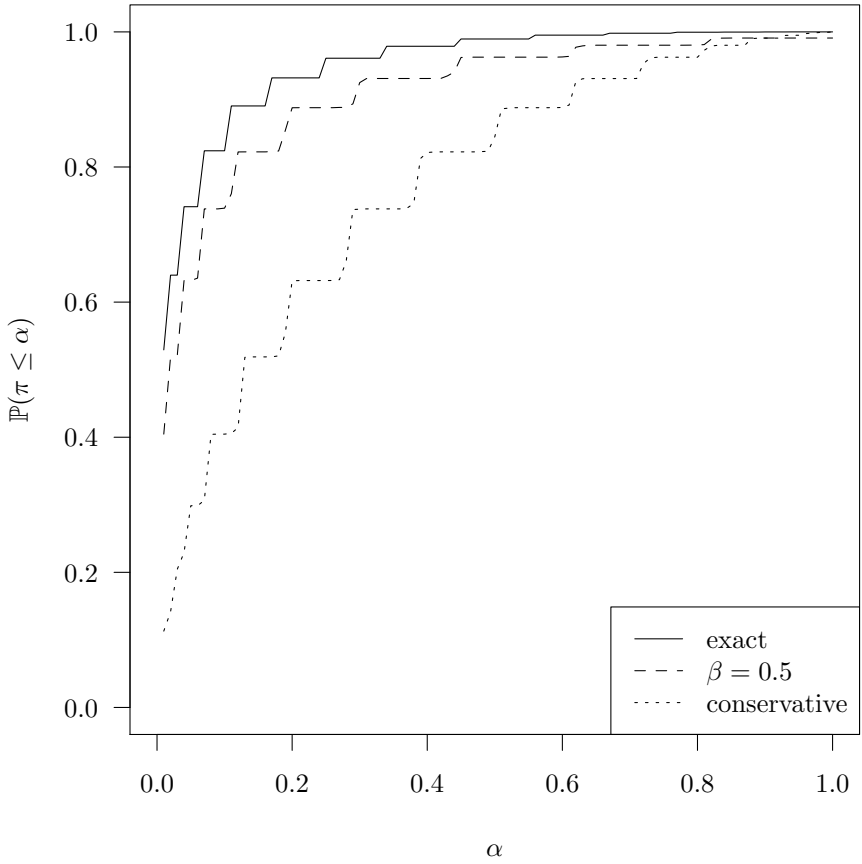

Figure 4.1.: Power for $\rho = 0.5$ in Example 4.5.

(Example 4.3) with $\gamma_0 = 0.5$ and 0.1. Figure 4.1 shows the power as function of α for the exact Monte-Carlo p-value, the conservative p-value and $Q_T(0.5)/0.5$ for $\rho = 0.5$.

The power of the de-randomized p-values is considerably better than that of the conservative. The choice of β has not a big influence on the power, therefore the adaptive version brings no benefit.

4.3.2. K-by-L Tables

Now suppose that we have multiple K-by-L tables $S^{(i)}$ with $\max(K, L) > 2$ and we want to test for independence. To get an exact p-value for one table $S^{(i)}$, we compute Pearson's chi-squared statistic $T_s^{(i)}$ for all t tables with the same marginal totals as $S^{(i)}$. The p-value is given by

$$\pi_i^{\text{cons}} := \frac{\#\{s \le t\colon T_s^{(i)} \ge T_0^{(i)}\}}{t},$$

where $T_0^{(i)}$ is the statistic of $S^{(i)}$. If t is small, π_i^{cons} can take only few values and therefore it would be worthwhile to consider the randomized p-value

$$\pi_i \sim \text{Unif}\left[\frac{\#\{s \le t\colon T_s^{(i)} > T_0^{(i)}\}}{t}, \frac{\#\{s \le t\colon T_s^{(i)} \ge T_0^{(i)}\}}{t}\right].$$

For multiple tables these p-values can be combined using (4.1). Alternatively, noting that $T_0^{(i)} \sim_{\text{appr.}} \chi^2_{(K-1)(L-1)}$, we define the combined p-value

$$\pi \;:=\; 1 - F_{m(K-1)(L-1)}\Big(\sum_{i=1}^{m} F^{-1}_{(K-1)(L-1)}(\pi_i)\Big), \tag{4.3}$$

where F_k denotes the c.d.f. of χ^2_k.

Example 4.6. Table 4.2 shows data from a hypothetical multi-center clinical trial. For each of the five 2-by-3 tables the p-value π_i^{cons} is given. Combining them with (4.1), we get $\pi^{\text{cons}} = 0.122$. If we combine the randomized p-values π_i using (4.1) and de-randomize the result, we end up with a considerably smaller p-value. For example $Q_T(0.1)/0.1 = 0.011$.

Combining the p-values with (4.3) leads to even better results, namely $\pi^{\text{cons}} = 0.051$ and $Q_T(0.1)/0.1 = 0.008$.

Table			p-value
0	3	3	0.3
3	0	0	
0	2	1	0.333
3	0	0	
0	0	3	0.5
2	1	1	
0	3	0	0.111
6	0	2	
0	0	2	0.333
3	4	1	

Table 4.2.: Data from a hypothetical multi-center clinical trial.

Appendix

A. Classical Results

A.1. Lindeberg-Feller Central Limit Theorem

Theorem A.1. *For $n = 1, 2, 3, \ldots$ let $\boldsymbol{Y}_{n,1}, \boldsymbol{Y}_{n,2} \ldots \boldsymbol{Y}_{n,n} \in \mathbb{R}^d$ independent random vectors. Suppose that for a matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$*

$$\sum_{i=1}^{n} \mathbb{E}(\boldsymbol{Y}_{n,i}) = \boldsymbol{0},$$

$$\sum_{i=1}^{n} \mathrm{Var}(\boldsymbol{Y}_{n,i}) \to \boldsymbol{\Sigma},$$

$$\sum_{i=1}^{n} \mathbb{E}\big(\|\boldsymbol{Y}_{n,i}\|^2 \min(1, \|\boldsymbol{Y}_{n,i}\|)\big) \to 0.$$

Then $\sum_{i=1}^{n} \boldsymbol{Y}_{n,i} \to_{\mathcal{L}} \mathcal{N}_d(\boldsymbol{0}, \boldsymbol{\Sigma})$.

This is a standard result in asymptotic statistics, see e.g. van der Vaart (1998).

A.2. Neyman-Pearson Lemma

Theorem A.2. *Let P_0 and P_1 be probability distributions possessing densities f_0 and f_1, respectively, with respect to a measure μ.*

(i) Existence. *For testing $H : P_0$ against the alternative $K : P_1$ there exists a test φ and constants $c \in [0, \infty]$ and $\gamma \in [0, 1]$ such that*

$$\mathbb{E}_0 \varphi(X) = \alpha \tag{A.1}$$

and

$$\varphi(x) = \begin{cases} 1 & \text{if } f_1(x) > c f_0(x) \\ \gamma & \text{if } f_1(x) = c f_0(x) \\ 0 & \text{if } f_1(x) < c f_0(x). \end{cases} \tag{A.2}$$

(ii) Sufficient condition for a most powerful test. *If a test satisfies* (A.1) *and* (A.2) *for some c and γ, then it is most powerful for testing P_0 against P_1 at level α.*

(ii) Necessary condition for the most powerful test. *If φ is most powerful at level α for testing P_0 against P_1, then for some c it satisfies*

$$\varphi(x) = \begin{cases} 1 & \text{if } f_1(x) > cf_0(x) \\ 0 & \text{if } f_1(x) < cf_0(x) \end{cases} \tag{A.3}$$

a.s. μ. It also satisfies (A.1) *unless there exists a test of size $< \alpha$ and with power* 1.

The proof can be found in Shao (2003).

References

Agresti, A. (2007). *An introduction to categorical data analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition.

Bache, K. and Lichman, M. (2013). UCI machine learning repository.

Bates, D. and Maechler, M. (2010). *Matrix: Sparse and Dense Matrix Classes and Methods*. R package version 0.999375-44.

Dümbgen, L. (1998). On Tyler's M-functional of scatter in high dimension. *Ann. Inst. Statist. Math.*, 50(3):471–491.

Dümbgen, L. (2010). *Empirische Prozesse*. Lecture Notes, University of Bern.

Dümbgen, L. (2011). *Multivariate Statistik*. Lecture Notes, University of Bern.

Dümbgen, L., Igl, B.-W., and Munk, A. (2008). P-values for classification. *Electron. J. Stat.*, 2:468–493.

Dümbgen, L., Pauly, M., and Schweizer, T. (2013). A survey of m-functionals of multivariate location and scatter. Technical Report 77, University of Bern. arXiv:1312.5594.

Friedman, J. H., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Kiefer, J. (1953). Sequential minimax search for a maximum. *Proceedings of the American Mathematical Society*, 4(3):pp. 502–506.

Kushmerick, N. (1999). Learning to remove internet advertisements. In *Proceedings of the third annual conference on Autonomous Agents*, AGENTS '99, pages 175–181, New York, NY, USA. ACM.

Lehmann, E. L. (1959). *Testing statistical hypotheses*. John Wiley & Sons Inc., New York.

McLachlan, G. J. (1992). *Discriminant analysis and statistical pattern recognition*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York. A Wiley-Interscience Publication.

Meinshausen, N., Meier, L., and Bühlmann, P. (2009). p-values for high-dimensional regression. *J. Amer. Statist. Assoc.*, 104(488):1671–1681.

Muirhead, R. J. (1982). *Aspects of multivariate statistical theory*. John Wiley & Sons Inc., New York. Wiley Series in Probability and Mathematical Statistics.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Shao, J. (2003). *Mathematical statistics*. Springer Texts in Statistics. Springer-Verlag, New York, second edition.

Taylor, A. E. and Lay, D. C. (1980). *Introduction to functional analysis*. John Wiley & Sons, New York-Chichester-Brisbane, second edition.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288.

van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York. With applications to statistics.

Zhu, J. and Hastie, T. (2004). Classification of gene microarrays by penalized logistic regression. *Biostatistics*, 5(14):427–443.

Zumbrunnen, N. (2009). P-values for weighted nearest-neighbor classifiers. Master's thesis, University of Bern.

Zumbrunnen, N. and Dümbgen, L. (2011). *pvclass: P-values for Classification*. R package version 1.01.

Index

List of Symbols

Symbols used in Chapters 1–3. This list is not exhaustive.

$B(\boldsymbol{x}, r)$	closed ball of radius r centered at $\boldsymbol{x}$, p. 7
$D_{\boldsymbol{\Sigma}}(\boldsymbol{x}, \boldsymbol{y})$	Mahalanobis distance between $\boldsymbol{x} \in \mathbb{R}$ and $\boldsymbol{y} \in \mathbb{R}$ with respect to $\boldsymbol{\Sigma}$, p. 4
F_ξ	distribution function of the one-dimensional random variable ξ, p. 48
$H(\boldsymbol{\beta}, \gamma)$	half-space in $\mathbb{R}^d$, p. 48
$I(\cdot)$	ordered elements of $\mathcal{G}_\theta$, p. 15
L	number of classes, p. 3
M	measure on $\mathcal{X}$, p. 3, 15
N_θ	number of training observations of class θ, p. 5, 15
P_θ	conditional distribution $\mathcal{L}(\boldsymbol{X} \mid Y = \theta)$, p. 3, 15
R_θ	missclassification rate, p. 9
R	risk of missclassification, p. 4
$S(\tau, \boldsymbol{X}, \theta)$	sum of test statistics, p. 39
$T_\theta(\boldsymbol{X}, \mathcal{D})$	test statistic based on training data, p. 16
$T_\theta^*(\boldsymbol{x})$	test statistic for the optimal p-value, p. 14, 15
$U(\boldsymbol{x}, r)$	open ball of radius r centered at $\boldsymbol{x}$, p. 7
$W_n(i)$	weight assigned to observation i, p. 9
Y	class label, p. 3, 15, 47
Z_1	first component of $\boldsymbol{Z}$, p. 47
$\mathbb{1}(A)$	indicator function for the set A, p. 4
$\boldsymbol{A}$	general matrix notation, $\boldsymbol{A} = \begin{pmatrix} A_{1,1} & \cdots & A_{1,d} \\ \vdots & \ddots & \vdots \\ A_{d,1} & \cdots & A_{d,d} \end{pmatrix}$, p. 48
$\boldsymbol{\Delta}_{\boldsymbol{A}}$	scaled difference $\sqrt{n}(\widehat{\boldsymbol{A}} - \boldsymbol{A})$, p. 48
$\boldsymbol{\Delta}_{\boldsymbol{v}}$	scaled difference $\sqrt{n}(\widehat{\boldsymbol{v}} - \boldsymbol{v}) = (\Delta_{\boldsymbol{v},1}, \ldots, \Delta_{\boldsymbol{v},d})^\top$, p. 48
$\mathbb{I}_d$	d-dimensional identity matrix, p. 47
$\mathcal{N}_d(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma})$	d-dimensional Gaussian distribution with mean vector $\boldsymbol{\mu}_\theta \in \mathbb{R}^d$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$, p. 4
P_0	distribution of $\boldsymbol{Z}$, p. 47
$\boldsymbol{\Sigma}$	positive definite covariance matrix in $\mathbb{R}^{d \times d}$, p. 4, 47
$\boldsymbol{U}^\perp$	orthogonal complement of $\boldsymbol{U}$, p. 47

$\boldsymbol{X}$	feature vector, p. 3, 15, 47
$\boldsymbol{Y}^{\boldsymbol{A}}_{n,i}$	summand of $\Delta_{\boldsymbol{A}}$, p. 48
$\boldsymbol{Y}^{\boldsymbol{v}}_{n,i}$	summand of $\Delta_{\boldsymbol{v}}$, p. 48
$\boldsymbol{Z}$	$\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_1 - \boldsymbol{\mu}_1)$, p. 47
$\boldsymbol{\beta}$	$\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$, such that $\|\boldsymbol{\beta}\| = D_{\boldsymbol{\Sigma}}(\boldsymbol{\mu}_2, \boldsymbol{\mu}_1)$; in the proof of Lemma 3.13 $\boldsymbol{\beta} := \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_\lambda - \boldsymbol{\mu}_\theta)$, p. 47, 73
$\mathbf{1}_d$	d-dimensional vector of ones, p. 31
$\boldsymbol{e}_1$	first standard unit vector, p. 50
$\|\cdot\|_\infty$	uniform norm, p. 55
$\|\cdot\|$	Euclidean or Frobenius norm for vectors or matrices, respectively, p. 5, 32
$\widehat{R}_\theta$	cross-validated missclassification rate, p. 9
$\widehat{\boldsymbol{A}}$	estimator for $\boldsymbol{A}$, p. 48
$\widehat{\boldsymbol{\Sigma}}_M$	M-estimator, p. 6
$\widehat{\boldsymbol{\Sigma}}_{sym}$	symmetrized M-estimator, p. 6
$\widehat{\boldsymbol{\Sigma}}$	standard estimator for $\boldsymbol{\Sigma}$ or one of the M-estimators $\widehat{\boldsymbol{\Sigma}}_M$ and $\widehat{\boldsymbol{\Sigma}}_{sym}$, p. 6
$\widehat{\boldsymbol{\mu}}_\theta$	standard estimator for $\boldsymbol{\mu}_\theta$, p. 6
$\widehat{\boldsymbol{v}}$	estimator for $\boldsymbol{v}$, p. 48
$\widehat{P}$	empirical measure of a sample of independent random variables with distribution P, p. 7, 54
$\widehat{R}(\boldsymbol{x}, \boldsymbol{X}_i)$	rank of training observation $\boldsymbol{X}_i$, p. 9
$\widehat{Y}(\boldsymbol{X})$	classifier, point predictor for Y, p. 4
$\widehat{Y}^*(\boldsymbol{X})$	optimal classifier, p. 4
$\widehat{f}_\theta$	estimator of f_θ, p. 5
$\widehat{r}_{k,n}(\boldsymbol{x}, \mathcal{D})$	radius of the smallest ball centered at $\boldsymbol{x}$, which covers at least k training vectors $\boldsymbol{X}_i$, p. 7
$\widehat{w}_\theta(\boldsymbol{x}, \mathcal{D})$	estimator of $w_\theta(\boldsymbol{x})$, p. 16
$\widehat{w}_\theta$	estimator of w_θ, p. 5, 47
$\widehat{\mathcal{I}}_\alpha(b, \theta)$	empirical conditional inclusion probability, p. 17, 70
$\widehat{\mathcal{P}}_\alpha(b, S)$	empirical pattern probability, p. 17
$\widehat{\mathcal{Y}}_\alpha(\boldsymbol{X})$	prediction region, p. 10
$\widehat{\mathcal{Y}}_\alpha(\boldsymbol{X}, \mathcal{D})$	prediction region based on training data, p. 15
$\mathbb{B}_{P,n}$	empirical process induced by $\mathcal{H}$, p. 54
$\boldsymbol{\mu}_{\theta,b}$	$(\boldsymbol{\mu}_\theta + \boldsymbol{\mu}_b)/2$, p. 5
$\boldsymbol{\mu}_\theta$	mean vector in $\mathbb{R}^d$, p. 4
$\otimes$	Kronecker product, p. 29, 48
$\pi_\theta(\boldsymbol{X}, \mathcal{D})$	p-value based on training data, p. 10, 16
$\pi_\theta(\boldsymbol{X}_i, \mathcal{D}_i)$	cross-validated p-value, p. 17

$\pi_\theta^*(\boldsymbol{x})$	optimal p-value, p. 13, 14
$\boldsymbol{\pi}$	vector of p-values $(\pi_\theta)_{\theta\in\mathcal{Y}}$, p. 13
$\triangle$	symmetric difference, p. 48
τ^*	optimal parameter τ, p. 40
$\tilde{\boldsymbol{X}}_i$	centered observation, p. 47
$\to_{\text{a.s.}}$	convergence in law, p. 48
$\to_p$	convergence in probability, p. 18, 48
$\to_{\mathcal{L}}$	convergence in law, p. 48
$\boldsymbol{u}$	$\|\boldsymbol{\beta}\|^{-1}\boldsymbol{\beta}$, p. 59
$\boldsymbol{v}_{i:j}$	vector consisting of the components i to j of $\boldsymbol{v}$, p. 48
$\boldsymbol{v}$	general vector notation, $\boldsymbol{v} = (v_1, v_2, \ldots, v_d)^\top$, p. 48
$\text{vec}(\boldsymbol{M})$	vector which is formed by stacking the columns of a matrix $\boldsymbol{M}$ (from left to right), p. 34, 48
$d(\cdot,\cdot)$	some metric, p. 7
$f(\boldsymbol{x})$	density of the random vector $\boldsymbol{X}$, p. 11
f_θ	density of P_θ with respect to M, p. 3, 15
$f_{\boldsymbol{\xi}}$	density of the random variable $\boldsymbol{\xi}$, p. 48
g_θ	continuous bounded function such that $g_\theta(r)r^2$ is bounded for $r \geq 0$, p. 56
h_θ	continuous bounded function, p. 56
n	training sample size, p. 15, 47
$w_\theta(\boldsymbol{x})$	posterior weight, p. 9, 14
w_θ	prior probability, p. 3
$w_{b,\theta}$	ratio of prior weights, p. 14
$z_{1-\alpha/2}$	$(1-\alpha)$-quantile of the standard Gaussian distribution, p. 71
$\mathcal{D}(\boldsymbol{X},\theta)$	training data extended by $(\boldsymbol{X},\theta)$, p. 21
$\mathcal{D}_i(\boldsymbol{X}_i,\boldsymbol{X},\theta)$	training data after adding the observation $(\boldsymbol{X},\theta)$ and with the class label of observation $\boldsymbol{X}_i$ set to θ, p. 39
$\mathcal{D}_i(\boldsymbol{x})$	training data with $\boldsymbol{x}$ in place of $\boldsymbol{X}_i$, p. 16
$\mathcal{D}_i$	training data without observation $(\boldsymbol{X}_i, Y_i)$, p. 9, 17
$\mathcal{D}$	training data, consisting of pairs $(\boldsymbol{X}_i, Y_i)$, for $i = 1,\ldots,n$, p. 5
$\mathcal{G}_\theta$	index set of training observations of class θ, p. 5
$\mathcal{H}$	collection of all half-spaces in $\mathbb{R}^d$, p. 48
$\mathcal{I}_\alpha(b,\theta \mid \mathcal{D})$	conditional inclusion probability, p. 17, 70
$\mathcal{L}(\boldsymbol{X})$	distribution of the random variable $\boldsymbol{X}$, p. 3
$\mathcal{P}_\alpha(b,S \mid \mathcal{D})$	pattern probability, p. 18
$\mathcal{R}_\alpha$	measure of risk for p-values, p. 13
$\mathcal{X}_0$	support of $\mathcal{L}(\boldsymbol{X})$, p. 7
$\mathcal{X}$	feature space, p. 3, 15

$\mathcal{Y}$	set of class labels $\{1, \ldots, L\}$, p. 3

Erklärung

gemäss Art. 28 Abs. 2 RSL 05

Name/Vorname:	Zumbrunnen Niki
Matrikelnummer:	04–124–269
Studiengang:	Statistik, Dissertation
Titel der Arbeit:	P-Values for Classification – Computational Aspects and Asymptotics
Leiter der Arbeit:	Prof. Dr. L. Dümbgen und Prof. Dr. A. Munk

Ich erkläre hiermit, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Mir ist bekannt, dass andernfalls der Senat gemäss Artikel 36 Absatz 1 Buchstabe r des Gesetzes vom 5. September 1996 über die Universität zum Entzug des auf Grund dieser Arbeit verliehenen Titels berechtigt ist.

Bern, 05. 03. 2014

Niki Zumbrunnen

Lebenslauf

1985	Geboren am 5. September in Bern
1992–1998	Primarschule in Bern
1998–2000	Sekundarschule in Bern
2000–2004	Mathematisch-Naturwissenschaftliches Gymnasium Bern-Neufeld (Schwerpunktfach Physik und angewandte Mathematik)
2004–2008	Bachelorstudium in Mathematik an der Universität Bern Minors: Philosophie und Naturwissenschaften Computational Science Informatik
2006–2009	Masterstudium in Mathematik an der Universität Bern
2009–2014	Doktorat an den Universitäten Bern und Göttingen